

Another anniversary to celebrate

Nature's pride in its survival this far should not be mistaken for evidence that the future can prudently be an extrapolation of the past. There are challenges and opportunities ahead.

BIRTHDAYS are usually ambiguous occasions; joy is modulated by the knowledge that time is passing inexorably. By contrast, the anniversaries that institutions enjoy can in principle be free from that ambivalence. There are academic institutions around the world that celebrate their birthday with an annual founders' day long after most people have forgotten who the founders were; that is celebration unalloyed. Periodicals such as this come somewhere in between. There is no natural term to their existence, but even when they have reasonable cause to be proud of their past, they know that the foundations of their future must be repeatedly relaid.

That is this periodical's mood on this 125th anniversary. The self-congratulation is the easy part. Among other things, *Nature* is proud of its origins, in the ferment caused in mid-Victorian Britain by the publication of Darwin's *Origin of Species by Natural Selection*, but even more so of the daring blend of journalism and intellectual preoccupation that sustained it through the remainder of the nineteenth century, when its contemporaries (including one with the interesting title *The New Age*) were dying off like flies. That this periodical became a scientific journal almost by accident has been amply remarked upon. That it became an international journal is again a legacy of the nineteenth century, and of the perceptiveness with which its founders cast their web of curiosity beyond Great Britain (as it was then called), not only over the rest of Europe but over the New World as well.

More recently, *Nature* has prospered as the research profession has grown. The past century has been a great adventure for all those concerned with science. It is hoped that the part of this issue (pages 11 to 36) labelled "Frontiers of ignorance" will suggest how much excitement (and reward) is still to come. Domestically, *Nature* has learned from

these developments of the need to scatter its editorial resources over the surface of the Earth. That process, begun in the past quarter of a century, will no doubt continue. Technology not only makes disseminated editorial function feasible, but offers many opportunities not yet exploited. It cannot be long before some of *Nature's* communication with its readers is electronic, for example.

But there is more to communication than the transfer of signals from one place to another. What there is to say also matters. In that connection, there is less to boast about. It is a matter for regret that so many communications in science are easily accessible only to those who happen already to know what they say. Mercifully, there is a chance that these problems will be more easily surmountable with the technology now becoming available. It will be excellent, especially for *Nature's* readers, if the result is a weekly putting of ink on paper in which the freshness and informality of the earliest issues can be harnessed to the communication of the intellectual content of contemporary research. That is one obvious goal to aim at.

The consequences of the competitiveness that now marks research within an enlarged community almost everywhere engaged in a competition for research funds is a matter with which the research community, not the journals, must come to grips. One consequence is the frequent publication of needlessly breathless research reports that are also insufficiently reflective. Another is an inevitable fragmentation of the literature of research. The importance of speed in the publication of interesting research is nowhere denied; others than its authors need to know, for science is a cumulative and collective enterprise. But the fragmentation is often an impediment to understanding.

A further challenge for the years ahead lies in the cultural diversity of the research enterprise. People work in different countries, often speaking (and writing) in languages other than English. It is notoriously the case that those preparing research reports refer most often to work by their compatriots, presumably because they are less well-informed about that of other people. Conversely, those whose first language is not English believe (often correctly) that they are at a disadvantage in the competition for space in international journals for that reason only. Even on the grounds of efficiency, there are strong grounds for breaking down those barriers. At the urging of a number of French researchers, *Nature* plans steps in those directions.

But efficiency is not, and should not be, the only motive.

Anniversary features

The following are items in this issue linked with this week's anniversary issue:

- The cover of this issue conceals a reproduction of a painting by Greg Griffiths specially commissioned for this issue.
- A brief note of the meeting in Paris last week jointly arranged by the Collège de France and *Nature* (see page 7).
- "Frontiers of Ignorance" (see pages 11–36).
- "Adjudication of prizes for News and Views contributions, page 37.



From the outset, *Nature* has been a kind of club in the best sense of that word. Its readers are also, potentially, its contributors. Many of its admitted idiosyncrasies are regularly forgiven by its familiars. Journalism, which at its best is a way (other than by research) of telling what the truth is, has been a crucial ingredient of *Nature*'s style from the outset, and will remain so. The hope must be that the opportunities presented by technology for the more rapid dissemination of information can also be used so as to strengthen the sense of intimacy between editors, contributors and readers that has been this journal's special claim on professional attention for the past 125 years. □

All change in Brussels

The European Commission in the New Year will influence European science, but nobody can yet tell how.

ONE of the quaint happenings in the European Union is that, from time to time, the old faces disappear, to be replaced by new ones. The next big upheaval will come in January, when several new people will come into office as members of the European Commission. As it happens, the commissioners responsible for science and industry are being changed. Most spectacularly, Signor Antonio Ruberti, who has managed the commission's research portfolio since the appointment of the last Italian government a year ago, is to be replaced by Mme Edith Cresson, who endured a turbulent brief spell in 1990 as the first woman to be the prime minister of France. It is also relevant and significant that Herr Martin Bangemann, a long-standing member of the commission and an adherent of the Free Democratic Party (FDP), the junior member of Germany's coalition government, becomes the commissioner responsible for industry, information technology and telecommunications; those are the fields in which much of the commission's research spending is concentrated.

So how will the new people shape up to their new responsibilities? Only guesswork can be a guide. But if Cresson feels like embarking on a scheme to do for Europe as a whole what the newly-elected French government did for France in 1980, she would have lots of people cheering for her. Unfortunately, it is more probable that she and Bangemann between them will be pushing for the continuation of the *status quo*, but for more of it. That would be a great misfortune. The trouble with Europe's research programme is that it has survived almost unchanged in its purpose since the first "framework" programme, which long predates the "Single Act" of 1986 that ushered in the now single market, not to mention the Maastricht Treaty, which speaks of the "coordination" of national research in Europe.

Will Cresson and Bangemann, between them, set about the radical review of Europe's research that the circumstances demand? That is where they should put their energy. And this is how they should set about the job. First, they should recognize that many of the principles on which existing projects are based have been overtaken by events. Can it still make sense, for example, to require that applica-

tions for projects in applied research, in information technology or telecommunications, should include partners from at least two and preferably more countries, when the single market decrees that commercial organizations be given every opportunity to drive each other into the ground? The commission would be sharply criticized if it abandoned the practice, but not on grounds of principle. National governments would simply complain that they and their constituents had been left out.

Second, Cresson and Bangemann should accept that the chief purpose of central spending on research by the commission will be, for the next ten years at least, the broadening and deepening of Europe's high-level technical skills. That argues, first, for supporting basic research as a priority, and applied research if there are funds left over. As things are, there are too many parts of Europe in which academic institutions are only indifferently engaged in contemporary science — and many others in which Europe's capacity to match insight with tangible (and perhaps patentable) discovery is comparatively impoverished. Cresson and Bangemann could do worse than begin by commissioning a study of what Europe's academic institutions have to offer in advanced skill — and then heeding the implications for improved high-level training in discovery and research.

Third, Cresson-Bangemann should look for a way of getting the bureaucracy off the back of European science. That is not to fall into the trap of suggesting that the Brussels bureaucracy is an army of overpaid international public servants with so much time on its hands that it can interfere in minor matters. On the contrary, the army is too small to discharge even routine tasks properly — announcing competitive awards of research funds in time for those likely to be interested to complete meaningful applications before the deadlines pass, for example. But the research community, which is well networked, is excellent at tasks of which the commission regularly makes a muddle. Cresson-Bangemann should resolve to delegate more of their evaluation to sympathetic external organizations, keeping to the commission the responsibility for strategy which it cannot dodge.

Cresson-Bangemann should also be careful to talk at length to Ruberti before he disappears. Unlike too many of his predecessors, Ruberti has not sought to put his stamp on the commission's research programme as a whole, and in too little time. Instead, he has been concerned with the infrastructure of research (the new data-centre at Seville is an example) and with trying to ensure that some things work a little better. That may not be much to boast about, but it is a lot better than the scant attention the commission's research arm has traditionally paid to the administration of research.

What Cresson-Bangemann should aim at is what Europe has been hungering for since the end of the Second World War: the resources with which to clothe its view of what the world is like with proof that the view corresponds with reality. The temptation, for Cresson-Bangemann, will be to look for yet another gimmick with which to fire the European imagination — an even faster train-set for adult passengers, for example. Far better to start by making a fair contribution to the understanding of the world we live in. □

NIH drops survey on discrimination, claiming questions 'ill-designed'

Washington. The US National Institutes of Health (NIH) have abandoned an equal-opportunities survey that had been sent out to each of their 16,000 staff members. Officials say the survey was stopped because employees complained about its design. But they admit that some returns were assessed before the decision was made to abandon the exercise.

The aborted study, planned to cost \$80,000, was intended to help the NIH Task Force on Fairness in Employment Practices to prepare an action plan to counter allegations of sexual and racial discrimination at NIH (see *Nature* 363, 105; 1993).

Staff were asked about unfair employment practices at NIH. But most of the 14

questions concerned the effectiveness of NIH's Equal Opportunities Office, whose manager has just been replaced.

Sandy Chamblee, acting deputy director for science policy at NIH and chairwoman of the task force, said that the task force decided to rescind the survey because it "didn't want to be side-tracked on the secondary issue" of the survey's validity.

The questionnaire was sent out on 1 July with a covering letter from Harold Varmus, the director of NIH, stating that summary results would be shared with all staff and asking for a response within five days. But after extensive debate within the task force, Varmus sent a second letter on 18 August saying that the survey would not proceed,

and that "no analysis will be conducted of the response received thus far".

The lateness of the rescission has led some NIH staff to suggest that the survey had been abandoned because the task force did not like the results. Anne Thomas, a spokeswoman for NIH, dismisses this suggestion as "ridiculous". But she concedes that the task force did indeed see "a small number" of the returns before deciding to abandon the survey.

NIH staff were asked to send the questionnaires to Research Applications, a Maryland market research company, whose fee was to be \$77,843. NIH said it "is now looking at how that price might come down since the analysis was not completed".

Members of the task force are being tight-lipped about the rescission decision, referring enquiries to Chamblee, who in turn chose to answer detailed questions through Thomas. Thomas said that objectors to the survey had made six specific criticisms.

For example, some objectors claim that the survey was biased, with some of the questions assuming that unfair treatment did exist at NIH, such as "What should NIH do to reduce unfair treatment?". It also directed answers by requiring respondents to pick from nine types of "unfair treatment", leaving no room for additional comment.

Some staff apparently objected to being forced to answer "yes" or "no" to certain questions, even though only one out of 14 — "Do you think you have been unfairly treated in your present job?" — required such an answer.

In addition, Thomas said that there were some "sweeping, but non-specific" complaints about the survey and others that "it didn't give the opportunity to reflect on reverse discrimination, where people feel they are discriminated against because they are not part of a minority group".

Some observers claim that this last complaint suggests the NIH may be experiencing a backlash against its equal opportunities work. Issues of both race and sex discrimination have recently been given a higher public profile than some managers and staff feel is appropriate.

But Naomi Churchill, who replaced Diane Armstrong as manager of the NIH Equal Opportunities Office in September, says she is not discouraged. "I haven't encountered much resistance," says Churchill. "Most people here want to do the right thing."

Earlier this month, Harold Varmus met congressional staff to discuss sexual discrimination at NIH. He told them that he did not much like the questionnaire and had sent it out "against his better judgement" before rescinding it.

Colin Macilwain

Tsunami experts left high and dry

Tokyo. As a result of a territorial dispute with Russia, the Japanese government has blocked Japanese specialists in tsunami — the large waves that follow earthquakes — from participating in an international survey of the Russian-held southern Kuril Islands.

The islands were hit by tsunami last month after a large earthquake off the coast of Japan's northern island of Hokkaido. Tsunami up to five metres high swept over the southern Kuril Islands after the earthquake struck on 4 October and several Russian fishermen lost their lives. Tsunami also hit Japan's Pacific coast, but the waves were only one to two metres high and caused little damage (see *Nature* 371, 549; 1994).

Communicating by electronic mail, tsunami specialists in Japan, Russia and the United States quickly laid plans for a survey of the islands. Y. Tsuji of the Earthquake Research Institute of Tokyo University was to have led a team of three Japanese scientists that would have joined at least four Russians and one US scientist in Yuzhno-Sakhalinsk in Sakhalin on 16 October, before proceeding to the islands.

Tsuji received an official invitation from Aleksey Ivaschenko, deputy director of the Russian Institute of Marine Geology and Geophysics, and says that everything was ready for obtaining visas for his team from the Russian Embassy in Tokyo. But when he contacted the Japanese Ministry of Education, Science and Culture to obtain official approval, he was told that the trip would be "illegal" and he and the other Japanese scientists "could lose their jobs".

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

After-effects of last month's earthquake.

The ministry told Tsuji that university scientists are government officials and are therefore banned from visiting the islands. Russia occupied the islands at the end of the Second World War and Japan has been lobbying hard in recent years to have them returned.

Apart from preventing Japanese participation in the survey, which Tsuji says has proceeded without them, the territorial dispute has made it difficult for the survey scientists to reach the islands. These are only a few tens of kilometres from the northern tip of Hokkaido and it would make a lot of sense for the survey team to approach them by ship or by air from Hokkaido.

But according to one of the survey participants, Harry Yeh of the University of Washington in the United States, this idea was never taken seriously because of the territorial dispute. Instead, the team had to be organized at the Russian Tsunami Warning Centre in Sakhalinsk, 400 km away from the islands.

David Swinbanks

Germany debates future of research centres

Munich. Many of Germany's 16 national research centres are to appoint more industrialists to their policy-making committees in response to a report commissioned by the research minister, Paul Krüger, published last June. The centres may also see core government funding for some of their programmes cut by half.

But the report, based on a year-long study by an industry-orientated committee chaired by Hartmut Weule, head of the research division of Daimler-Benz, has opened up a fierce debate about how much say industry should have in controlling the future directions of national research efforts.

The report focused in particular on the activities of two of Germany's largest national research centres, the Forschungszentrum Jülich, known as the KfA, and the Kernforschungszentrum Karlsruhe, or KfK.

These two centres were established in the 1950s to develop Germany's nuclear energy technology, and have since evolved into interdisciplinary research centres. But along with the other national research centres, which absorb 70 per cent of the research ministry's budget, they have recently come under attack for the alleged inflexibility of their research programmes (see *Nature* 363, 292; 1993).

The report made a series of recommendations on how the centres could become more useful to industry, the most controver-

sial of which was the suggestion that the proportion of applied research carried out at the two centres should be increased from 30 per cent to 75 per cent within five years.

This caused an immediate outcry from the centres themselves, as well as from Germany's major research organizations such as the Max Planck Society and the Deutsche Forschungsgemeinschaft (DFG), which funds university research projects.

These organizations were relieved that the recommendation was rejected by the ministry of research as "unrealistic". But the ministry is taking two other recommendations of the report more seriously. These concern the restructuring of the centres' research-policy-making committees and the way their funds are allocated.

The ministry has asked all national research centres that carry out a significant amount of applied research to increase the proportion of industrialists on their scientific advisory committees to around 50 per cent from next year. These committees, which advise the decision-making supervisory boards of the national research centres, tend to consist of academics in the main.

The ministry has also said that there should be an increase in representation of industry — from around a quarter to around a third — on the supervisory boards themselves. How much extra control this will give industry over the direction of research is not yet

clear. Satisfaction with the system depends on two possibly conflicting factors. All agree that industrial representation should be of a high quality. But the ministry spokesman, Manfred Dietrich, says that to ensure that the best people from industry are prepared to devote time to the centres, they should be confident that their recommendations will have a strong chance of being adopted.

But Joachim Treusch, the director of KfA, who has already created slots in his centre's advisory subcommittee for industrial representatives, says that industry's needs should not dominate the research agenda.

Paul Krüger, who remains acting research minister until Germany's new cabinet is announced later this month, insists that no programme will be imposed on the centres by industry without the general support of their research scientists.

The ministry is also debating the report's recommendation that up to half of its core funding for some programmes at the centres, particularly those identified as being close to the market, should be cut, to be made up through competitive grants or industrial contracts.

The questions of who should decide which projects fall into this category, and where scientists should turn for the top-up funds, are currently under discussion. But Dietrich confirms that the ministry wants the change to be implemented within four years.

With no sign of more money on the horizon, Germany's scientific community is worried that throwing national research centres into competition for ever-tightening research grants and shrinking industrial research funds will be damaging.

Karl-Heinz Hoffmann, head of Germany's science council, the Wissenschaftsrat, which advises the government on research issues, says that a move from core funding to a reliance on competitive grants from government-funded research programmes would be "unthinkable" without an increase in general funding for such programmes.

The two new national research centres in the new *Länder*, for molecular medicine at Berlin-Buch and for environmental research at Leipzig, were both intended to rely more heavily on competitive grants than their western counterparts. But they have not been as successful in attracting money as predicted, he says, as there is not enough to go round.

How the discussions develop will depend on the shape of the new government. The so-called 'holy alliance' — the Max Planck Society, the DFG, the national research centres, the university rectors conference, the Fraunhofer Society and the Wissenschaftsrat — are meeting Krüger this week to discuss the research policy of the next government. The policy of the research centres will be at the top of their agenda.

Alison Abbott

UK labs to stay in public ownership

London. David Hunt, Britain's minister for science, announced last week that the government had decided against 'privatizing' the Daresbury and Rutherford Appleton Laboratories (DRAL), the main central research facilities for UK universities.

The two laboratories, jointly run by the Science and Engineering Research Council until 1 April of this year, will become an independent body operating within the Office of Public Service and Science.

"Daresbury and Rutherford are national assets," Hunt said in announcing his decision. "I believe that according them independent status is the right way to proceed, but with strong emphasis on operating in a commercial manner, and particularly meeting the needs of customers."

The future of the DRAL had been up in the air ever since the government decided to carry out a study of whether it should remain a public body. Earlier this year, Britain's Royal Society, together with a number of other bodies, warned the government against any effort to turn the joint organization of the two laboratories into a private institution (see *Nature*, 368, 382; 1994).

Their message was reinforced by a report prepared by consultants KPMG, which told the government that the services provided by DRAL were widely appreciated by the scientific community, both in Britain and overseas, and that privatization would not represent a sensible option. It suggested that DRAL should become a "non-departmental public body".

The government now appears to have accepted this advice. Its move was welcomed by the Royal Society, which issued a statement emphasizing that the government needs to ensure that DRAL continues to provide "vital support for members of the scientific community, irrespective of their sources of funding."

Under the new arrangements, the chief executive of the DRAL will have the same reporting relationship to the Director General of Research Councils as the heads of the six research councils. The budget of DRAL will be made up of a combination of core funding and money received directly from its 'customers'. The precise balance between the two has yet to be determined.

David Dickson

Frank Press to head new inquiry into US science policy

Washington. Frank Press, former president of the National Academy of Sciences and before that President Jimmy Carter's science adviser, is to lead a major inquiry into how the federal government allocates research and development (R&D) money in the United States.

The study will begin next month and is due to be completed in December 1995. According to Press, it will consider all federal R&D spending — which totalled \$73 billion this year — except for the large amounts spent on the development and trials of military equipment.

The report was requested by the Senate's labour, health and human services and education appropriations subcommittee, chaired by Tom Harkin (Democrat, Iowa), in a recent budget bill. Press says that its scope will be more comprehensive than any other undertaken in recent years.

"It is an extraordinary opportunity," he says, adding that the exercise "could have the impact of a Vannevar Bush report", a reference to the influential report *Science: The Endless Frontier* that laid the basis of US science policy after the Second World War.

Members of the panel are being selected, and will be announced next month. The panel will function under the auspices of the National Research Council, the research arm of the National Academies of Science and Engineering and the Institute of Medicine.

The Senate subcommittee, which is providing \$750,000 to pay for the study, said it was worried that funding for medical research is growing more slowly than the rate of inflation. "The study should consider the criteria that should be used in judging the appropriate allocation of funds to research and development activities," it said.

The work of the panel is likely to have a big impact not just on biomedical research — Harkin's main area of interest — but across the entire federal R&D programme. Many policy-makers feel that the priorities of the vast programme have been left badly out of date by the ending of the Cold War.

A recent OSTP science policy report — *Science in the National Interest* (see *Nature* 370, 5; 1994) — made the case for basic science. But it did not deal with the more difficult area of setting new R&D priorities for the nation.

Colin Macilwain



Press: 'Could have impact of Bush report'

US to use science links to meet foreign policy goals

Washington. President Bill Clinton's administration has selected six countries — Russia, China, India, South Africa, Argentina and Brazil — with which it will seek to secure foreign policy goals through collaboration in science and technology. One possible model is the recent agreement to involve Russia in the international space station.

Jane Wales, assistant director for national security and international collaboration at the White House Office of Science and Technology Policy (OSTP), revealed the strategy to an advisory panel of scientists and industrialists, which met for the first time in Washington last week.

Some members of the panel — the President's Council of Advisors on Science and Technology (PCAST) — expressed reservations about the durability of such collaborations and their impact on US employment.

Wales said that each of the six countries was chosen on three criteria: that each had had a real or potential nuclear capability; that each was going through a period of political transition; and that each offered major potential markets for US exports. "We believe we can apply science and technology collaboration to building stability" in these countries, she said.

According to Wales, the administration has already conducted fact-finding missions to four of the six countries. But further details of the approach remain vague. The international space station collaboration with Russia is one model, officials say, and a science and technology collaboration deal due to be signed with China next month will provide another.

Officials concede that building public support and obtaining funds from Congress for such collaborations will not be easy. David Shaw, for example, a member of PCAST who helps finance technology-based companies, said that the approach was likely to "result in the export of a large number of jobs of the very kind we most want to keep".

Charles Vest, president of the Massachusetts Institute of Technology (MIT) — citing the destruction of MIT's work with Japan before the Second World War — suggested that US institutions may be reluctant to work too closely, or too publicly, with some of the countries on the list.

The first meeting of PCAST, long-delayed since its announcement more than a year ago, had some difficulty in establishing a coherent agenda. After a wide-ranging discussion, the panel, which is co-chaired by John Young, former president of Hewlett-Packard, and Jack Gibbons, science adviser to Clinton, agreed to confine itself chiefly to reviewing strategy documents produced by the nine subpanels of the administration's National Science and Technology Council.

Vice president Al Gore met the panel, before its meeting, but left no clear message of what he wanted its members to do.

David Beckler, a former White House official with extensive experience of previous such panels, told the meeting that "receptivity" — an assurance that the president wanted your advice — was a requisite for success, quoting for example the need to provide information about surprise nuclear attacks. Without such receptivity, he warned, the work of such panels was "just an intellectual tennis game".

Colin Macilwain

Scientists support expelled researcher

Paris. Over 4,000 French scientists have signed a petition in support of Abderrahmane Bahri, an Algerian national and an engineer at the Centre National de la Recherche Scientifique (CNRS), who was imprisoned in Folembay army barracks in August and later expelled from France for allegedly having links with terrorist groups in Algeria.

Charles Pasqua, the minister of the interior, expelled Bahri using the "absolute emergency act" which allows the government to bypass the usual judicial procedures on the ground of state security.

The expulsion order accuses Bahri of "actively supporting in Europe the forwarding and distribution of material destined for armed groups in Algeria". Bahri, it says, has "numerous links" with senior

members of terrorist organizations, and posed a "threat to national security and public safety".

But Bahri's colleagues at the CNRS National Synchrotron Facility (LURE) at Orsay near Paris, have formed a "committee of solidarity" to secure both his return to France, and the right to answer the charges against him in a normal judicial hearing.

Both the board of the University of Paris XI at Orsay and the 'Comité Paritaire' of CNRS have passed similar motions, while the 'Comité de Défense des Hommes de Science' of the French Academy of Sciences — chaired by the Nobel prizewinner François Jacob — and the Human Rights Commission of the French Physical Society have written to Pasqua with the same request.

Declan Butler

Institute to boost Japan sequencing efforts

Tokyo. Japan's first institute dedicated to the sequencing and analysis of DNA was officially opened last week. The Kazusa DNA Research Institute is unusual in that it is financed by the local prefectural government rather than the central government. Its large-scale sequencing facility will be open to outside users and visiting researchers from Japan and overseas.

The institute, located in a new science research park in the Chiba Prefecture east of Tokyo, cost ¥15 billion (US\$150 million) to build and is equipped with a high-powered computer system linked into international and domestic computer networks. It has a large facility for DNA sequencing run by specially trained technicians.

Scientists at the institute, who are drawn from both universities and industry, will initially focus on two projects: the sequencing and analysis of the whole genome of a blue-green algae (cyanobacterium) and the

sequencing of large (more than 2 kb) full-length human cDNAs.

Plans for the institute were announced in early 1991 (see *Nature* **349**, 640; 1991), and the past three years have been spent training a team of 15 technicians to operate workstations and sequencers and to use computers to assemble sequences. There is a similar number of ordinary laboratory technicians, so the technical staff outnumber the current complement of 26 researchers.

According to Mitsuru Takanami, the president of the institute, the technicians had to be trained from scratch because such individuals are rare in Japan's public sector research system. Many more will be needed before the facility reaches full operational capability.

The institute has also had difficulty recruiting scientists. "Generally speaking, researchers in Japan like to hold more stable positions in national universities and com-



Mixed blessing: the Kazusa's countryside setting has drawbacks for house-hunters.

panies," says Takanami. Housing has posed a particular problem — it is not easy to buy a house in Japan and people are reluctant to move, particularly to the countryside.

A special floor in the institute has been set aside for joint research with public institutions and private companies. At present, it has several well-known "guest laboratory directors", including Michio Oishi, director of the Molecular Cellular Biology Laboratory of Tokyo University and Minoru Kanehisa of the Chemical Research Laboratory of Kyoto University who also heads Japan's Human Genome Centre, a genome database facility at Tokyo University (see *Nature* **361**, 484; 1993).

The laboratory for sequencing the cyanobacterium genome plans to start by determining the whole genome sequence and then work on the structure and function of photosynthetic genes. This genome was chosen because of its manageable size (3.8 Mb), and because of the group's interest in elucidating the photosynthetic system.

A second group is working on large full-length human cDNAs. Takanami says such long cDNA is being neglected by human genome researchers around the world because short cDNA sequences are much easier to handle. So far, in preliminary work, the group has sequenced 100 such long full-length cDNAs with a total length of 0.4 Mb.

Another aim of the facility is to develop DNA sequencing technology. Japan's attempts to develop efficient large-scale automatic sequencers have not so far been particularly successful. The Science and Technology Agency (STA) for several years funded a project to develop components of an automatic sequencing system in various companies.

But attempts to put the various machines together into one system at the Tsukuba Life Science Center of the Institute of Physical and Chemical Research have not been very successful and Japanese genome researchers still rely heavily on US sequencing technology. Takanami hopes his institute will alter this situation.

David Swinbanks

Euro-lobby backs animal research

Munich. More than 400 academic and industrial scientists are meeting in Strasbourg next week to inaugurate a new European-wide association to defend the use of laboratory animals in biomedical research.

The idea for the European Biomedical Research Association (EBRA) was first discussed at a meeting of the European Neuroscience Association in Madrid last year, in response to the activities of animal rights' groups. The founding members represent 25 European countries.

Ivar Aune, secretary of the Society for Health and Research — Germany's main research lobby group — points out that critics of the use of animals in research now wield considerable influence with the European Parliament. He argues that the benefits that animal experimentation bring to society, particularly in medicine, are often overlooked by such critics and their supporters.

EBRA will act as a representative voice for national research lobby groups. Ray Guillery, professor of human anatomy at the University of Oxford and chairman of the association's planning committee, says EBRA plans to educate the public, particularly schoolchildren, about the importance and necessity of animal research.

The group also aims to become an influential voice in legislation affecting animal research, particularly at the level of the European Union.

The association hopes to receive financing primarily through the corporate memberships of pharmaceutical companies and scientific societies, which will also provide a broad membership base.

Many research scientists, increasingly alarmed at the force of protests against ani-

mal experiments, have welcomed the initiative. Norman Bowery, professor of neuropharmacology at London's School of Pharmacy, particularly applauds EBRA's aim of improving laboratory animal welfare in member countries with less stringent controls.

But not everyone is convinced of the need for another group to defend research. Sir Mark Richmond, for example, head of research and development for Glaxo, is worried about duplication of effort between national research lobbies and EBRA.

Richmond is reserving judgement on whether Glaxo, which is contributing to the costs of next week's meeting, should provide the association with more permanent support. "We will have to see how the situation evolves," he says.

Others, such as Dirk Bootma, a geneticist from the University of Rotterdam, argue that the strongest efforts to defend research should continue to take place at the national level. But national lobby groups are very enthusiastic about increasing European-level collaboration. Mark Matfield, head of the United Kingdom's Research Defence Society (RDS), and also a prominent member of EBRA's planning committee, says that the two most active national groups — the RDS and the German association — are working closely with the founding committee. The RDS, for example, will offer office space and services for the new body.

Priorities for the future direction and structure of EBRA will be decided at the inaugural meeting next week. High on the list is whether to establish a joint policy on primate research, an area heavily targeted by protesters.

Toni Feder

Britain backs off compulsory year prior to PhD work

London. The British government formally announced last week the introduction of a new one-year Master's qualification, to be known as the MRes, intended to prepare graduates before embarking on a three-year PhD. But, to the relief of many in the scientific community, it said that the new qualification will not be a pre-requisite for PhD students supported by the research councils.

The Office of Science and Technology (OST) is inviting the research councils to run a series of pilot courses from autumn 1995. Their aim will be to provide potential PhD students with an introduction to research techniques — and to introduce a new qualification that will be of value to employers in its own right for those who decide not to pursue a PhD.

But, in announcing the government's decision, David Hunt, the minister for science, said that there would be "no requirement" for students to complete a Master's year before embarking on a three-year PhD. This is a significant change from the position of the government in last year's white paper, *Realising Our Potential*, which stated that taking a Master's degree would be the "normal pattern" for most PhD students.

At the time, the proposals received a mixed welcome from industry and academic institutions. Among concerns raised were that, without an increase in funding, a switch from three- to four-year postgraduate studies would inevitably lead to a significant

reduction in the number of PhDs produced by British universities.

Some of these concerns were addressed in a consultation document on the proposals published by the OST in February this year (see *Nature* 367, 674; 1994). This proposed that the MRes could be awarded to those completing the first year of a four-year PhD programme that would incorporate the aspects of the MRes in the first year of study.

Paul Leonard, executive for science and technology of the Chemical Industries Association, which had argued strongly against a compulsory one-year pre-PhD course, says that the turnaround should not be seen as a sign of weakness on the part of the government. "Changing one's mind means taking advice and acting accordingly," he says.

The government's decision was also welcomed by the Royal Society, which had expressed reservations about the initial proposals — and had suggested that a pilot scheme would be a good way of testing their viability. But the society is not totally convinced that the threat to PhD numbers has been lifted.

Joe Vinen, professor of physics at the University of Birmingham and chairman of a committee set up by the Royal Society to report on the proposed MRes, said he was pleased to see the government had incorporated flexibility. But, he added, "we would continue to be concerned about numbers of PhDs."

Maggie Verrall

Meeting celebrates 125th birthday

Paris. French President François Mitterrand last week expressed his support for a collaboration between *Nature* and the Collège de France in Paris to launch jointly an annual event on a topic of contemporary scientific interest.

His message was addressed to a colloquium held in Paris last week-end and jointly organized by the two bodies as the first of a series of events being mounted to celebrate *Nature's* 125th anniversary.

The colloquium provided an opportunity for several of the college's most distinguished scientists to reflect in public on the principal events in their academic careers. The biologist François Jacob, for example, spoke of the events that had led him to collaborate with fellow countryman Jacques Monod, and to later collaboration with US and British colleagues.

Physicist Anatole Abragam described his discovery of the English word "serendipity" as a description of some of his experiences with nuclear-magnetic resonance during a two-year stay at the University of Oxford in

the 1940s. Similarly Christian de Duve pointed out how many of his discoveries in the field of cell biology would not have taken place if those providing him with financial support had insisted that he focused solely on the mechanism of insulin, the topic for which he was being funded.

Etienne Baulieu drew on his experiences with the controversies surrounding his 'abortion pill' RU486 to make a plea for scientists to become more involved in public debates on the impacts of their discoveries. Palaeoanthropologist Yves Coppens described how his predecessors had refused to accept Neanderthal man as an ancestor "because he was so ugly".

Astrophysicist Jean-Claude Pecker, after describing the various pleasures he had experienced sky-watching from observatories around the world, made a plea for greater support for scientists wishing to publish in languages other than English. Other speakers were Jean Dausset, founder of the Centre des Etudes du Polymorphisme Humaine (CEPH), and the biophysicist Pierre Joliot. □

New EC appointees 'will seek to bind research to industry'

Paris. The joint research programmes of the European Union (EU) are likely to place greater emphasis on focused industrial and social goals following the confirmation last weekend that Edith Cresson, the former French socialist prime minister, will succeed Antonio Ruberti as the EU's research commissioner when he steps down at the end of the year (see *Nature* 371, 728; 1994).

As prime minister, Cresson was a strong advocate of an interventionist industrial policy. This has led to speculation that her appointment will lead to closer links between the directorate for research and development (DGXIII) and the directorates for industry (DGIII) and information technology (DGIII) (see *Nature* 371, 190; 1994).

Cresson has not, as some had expected, taken direct control of several of DGIII's current functions, such as competitiveness. She has only obtained direct control of the SPRINT and VALUE innovation and technology transfer programmes, which are currently run by DGXIII.

But at a meeting earlier this week, Cresson and Martin Bangemann — the German commissioner who will retain responsibility for both industry and information technology in the new commission — agreed to "take all useful and appropriate initiatives to launch common projects of industrial interest".

The fourth Framework programme, for 1994–98, has a budget of Ecu12.3 billion (US\$14.9 billion). One commission official says that Cresson will probably have a "margin of manoeuvre" to give the programme an "industrial flavour", using Ecu700 million held in reserve for the next Framework, and Ecu1 billion of anticipated contributions from the four countries (Austria, Finland, Sweden and Norway) which are expected to join the EU.

Commission officials say that Cresson's role will be reinforced by the expected appointment of François Lamouroux as her senior official. Lamouroux was deputy head of Cresson's office during her spell as prime minister, and is now a powerful official in the industry directorate. "He is a hard man, and a tough negotiator", says one official.

Another appointment announced last weekend which is likely to have a significant impact on areas of research such as biotechnology is that of Ritt Bjerregaard, a socialist member of the Danish parliament, as commissioner for the environment (DGXI). The appointment has already been welcomed by environmentalist groups. "We are glad to have a Dane," says Linda Bullard, from the green group at the European Parliament. "They are more environmentally friendly than other member states."

Declan Butler

Republican gains could mean tighter budgets for science

Washington. With the US congressional elections less than a week away, the Republican party stands its best chance since the days of Ronald Reagan of winning a majority in at least one chamber. And, if Republicans follow through on campaign promises to cut government spending, the result is likely to be less federal support for science.

In the 100-member Senate, Republicans only need an additional seven seats to seize the majority from the Democratic party. A Republican majority in the 435-member House of Representatives, which Democrats have controlled since the 1950s, is considered more of a long shot. The Republicans need 40 seats; but that could still happen.

A big Republican win would shift power in congressional committees, whose chairs go automatically to the majority party.

For example, that of the subcommittee that sets budgets for agencies such as the National Science Foundation (NSF) and the National Aeronautics and Space Administration (NASA) would probably go to Phil Gramm (Republican, Texas).

Gramm is a leading conservative and foe of President Bill Clinton. He favours space spending; like the current Democrat chair, Barbara Mikulski of Maryland, his state includes a major NASA centre. But he is also a hawk on the deficit.

Pete Domenici of New Mexico would be a leading contender to head the Senate committee that oversees energy research. If so, the two Department of Energy laboratories in his state — Sandia and Los Alamos — could expect to fare well.

Biomedical research is likely to suffer under Republican leadership. The Senate committee that oversees such research might pay closer attention to 'hot' topics for conservatives, such as human embryo research, even though the likely chair, Nancy Kassebaum of Kansas, is a moderate.

In the House of Representatives, George Brown (Democrat, California), the chairman of the Committee on Science, Space and Technology, is running an extremely close race against his Republican opponent. If he loses, science will be without one of its staunchest allies in the next Congress.

If Brown wins, but the Republicans gain a majority in the House, Robert Walker of Pennsylvania would be a candidate for his chair. But Walker might alternatively take

over the powerful Republican Whip position being vacated by his friend Newt Gingrich of Georgia.

In either scenario, Walker could become an important figure for science in the next Congress. In the past he has advocated consolidating federal research into a single cabinet-level department of science and technology. He has also led moves to reduce energy research and cut overhead payments for universities conducting government-funded research.

In a paper outlining his plans if Republicans were to take over the House Science Committee, Walker calls for strict fiscal discipline on science and technology projects, with the implementation of "absolute funding caps, and strict time and completion specifications". He would also continue Brown's fight against 'earmarking' funds for non-peer-reviewed projects.

Overall, however, changes at the committee level are likely to have only a marginal impact. The panels that oversee federal research tend to be less partisan than other congressional committees. Besides, there will be fewer dollars to argue about — and the situation will become worse if Republicans and reform-minded Democrats become serious about reducing the national debt. Trimming the size of government is a campaign theme that has done well with voters this year.

Gingrich and more than 300 fellow Republicans — a majority of those running this year — made a public show in September of signing a *Contract with America* that promises, among other things, to introduce a constitutional amendment requiring the government to balance its books every year.

Last year's US budget deficit was about \$200 billion so a balanced budget amendment would have profound effects. Before cutting into 'entitlement' programmes such as social security, Congress is likely to wring every last penny from 'discretionary' programmes such as research.

The *Contract with America* in fact identifies specific examples of possible budget 'offsets' in the area of science. These include limiting the growth of the NSF, abolishing the US Geological Survey and the new National Biological Survey and killing Clinton's showcase for government-industry partnership, the Advanced Technology Program (see right).

Such suggestions have surfaced in earlier Republican budget-cutting plans and even a Republican staff member for the committee that wrote them admits "some are not terribly well thought out yet". But they indicate where a Republican-led Congress might go searching for cuts.

Tony Reichhardt



Walker: could become key player.

US to fund project to develop desk-top DNA screening tool

San Francisco. A consortium led by the biotechnology company Affymetrix Inc. and including research groups from four West Coast universities has won one of the largest advanced technology grants being offered by the US Department of Commerce's National Institute of Standards and Technology (NIST).

The grant of \$31.5 million over five years will be used to support the development of a hand-held DNA diagnostic device that is sufficiently simple and compact to be used in physicians' surgeries.

The device would use the company's 'DNA chip', a silicon wafer containing thousands of DNA probes, which Affymetrix initially developed to detect drug-resistance mutations in HIV, the virus that causes AIDS.

The consortium plans to develop the DNA chip and integrate it into a single device to carry out DNA extraction, amplification and analysis. The chip could contain probes for any known genetic mutations, offering a battery of DNA diagnostics.

David Singer, president and chief executive of Affymetrix, which is based in Santa Clara, California, says that as the Human Genome Project rapidly generates complex information about the genetic basis of human diseases, the demand is growing for cost-effective tools able to apply such knowledge in the clinic.

Of the grant to the Affymetrix consortium, \$10 million will go to the company Molecular Dynamics to develop its capillary-array electrophoresis technology for separating and measuring DNA fragments. Lawrence Livermore Laboratory will provide technology for a miniaturized PCR amplifier the size of a Walkman radio.

According to Singer, the University of California, Berkeley, and the California Institute of Technology in Pasadena will both contribute expertise in high sensitivity dyes, while the University of Washington in Seattle and Stanford University will help to test the system as a sequencing tool.

The grant specifies that all intellectual property resulting from the collaboration will go to the company. "Universities don't like it," admits Singer. "But it ensures that patenting or ownership negotiations don't make it difficult to get together."

Other NIST Advanced Technology Program grant winners in the DNA diagnostic area included Vysis, an Illinois company also developing a DNA chip; E. I. du Pont de Nemours & Co. of Delaware, which is working on a rapid DNA diagnostic system to detect contamination in food; and Molecular Tool of Maryland, which is creating a simple, compact DNA-typing instrument.

Sally Lehrmann

Compromise sought over germplasm access

Washington. A new set of rules covering the potential use by agricultural biotechnology companies of germplasm stored in international agricultural research centres (IARCs) across the world has been proposed by the director general of the International Plant Genetic Resources Institute (IPGRI).

The proposed rules would require private-sector recipients of germplasm held by the gene banks to agree not to seek intellectual property protection on such material unless it was substantially modified through breeding programmes.

At the same time, companies wishing to market agricultural products derived from the germplasm would be required to negotiate royalty agreements with the countries from which it originally came.

The proposals were presented to the annual meeting of the Consultative Group on International Agricultural Research (CGIAR) at the World Bank in Washington last week. CGIAR is a confederation of 16 IARCs spread across the globe and widely recognized as having brought green revolution agriculture to the developing world.

Geoffrey Hawtin, director general of IPGRI, which is based in Rome, told the meeting that a new environmental ethic, based on the concept of the ownership of genetic resources and promoted by the United Nations (UN) Convention on Biological Diversity agreed in Rio de Janeiro in 1992, threatened to deter future increases in world food production by restricting the free flow of germplasm.

Supporters of the ownership concept argue that it will provide financial incentives for the conservation of biodiversity. But ironically, said Hawtin, it could wreak havoc with CGIAR breeding programmes, as these can require parental lines from more than half a dozen countries to produce new high-yielding varieties of rice, wheat or maize.

The CGIAR routinely releases these new varieties free of charge to national breeding programmes. Hawtin's proposal would require standardized material transfer agreements to accompany every shipment of germplasm from source countries to CGIAR gene banks.

Under the new proposals, "if you commercialize something, everyone, both North and South, will share the benefits", says Hawtin. "It's never going to be perfect, but frankly the alternative is that no materials move [between breeding programmes]."

At last week's meeting, the CGIAR also signed a memorandum of understanding with the UN Food and Agriculture Organization (FAO) placing nearly 500,000 genetic accessions within its gene banks "in trust" under the UN.

The move to name the FAO as the trustee of CGIAR germplasm is intended to fore-

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

A 'gene fund' could help pay for gene conservation programmes

stall accusations by source countries that either the CGIAR or the World Bank now intends to lay claim to these valuable genetic resources. Under the terms of the new memorandum, the CGIAR promises not to seek intellectual property rights on any germplasm held in its collections.

Although IPGRI is known for promoting *ex situ* conservation of germplasm around the globe, there is growing pressure on the CGIAR to devise new incentives for the *in situ* preservation of local or traditional crops, known as landraces.

M. S. Swaminathan, director of the Centre for Research on Sustainable Agriculture

and Rural Development in Madras, India, submitted a report containing a proposal for a system-wide intellectual property policy. He suggested that one way of doing this would be to introduce the concept of "farmers' rights" to protect the intellectual property in informal agricultural innovations.

Swaminathan told the CGIAR meeting that farmers' rights protection — like patents or plant-variety protection — would allow farm-based innovators to profit from their inventiveness and their efforts to conserve landraces *in situ*.

Brushing aside criticism of the practical obstacles facing such a scheme, Swaminathan pointed to the current debate in India on draft legislation to create a system of "genetic identity centres" able to establish the ownership of informal genetic innovations with local crops.

If such landraces were eventually bred into hybrid commercial varieties, then the fact that they had been registered with the Indian Ministry of Agriculture would create the basis for claiming royalties on their use.

In order to simplify the thousands of competing claims that would result from such a system of farmers' rights, Swaminathan and others have proposed the creation of a "community gene fund", financed through a five per cent royalty or tax on the gross sales of commercial hybrid seeds derived from registered landraces.

Such a gene fund, he argued, could pay for hundreds of small-scale *in situ* conservation projects throughout India, helping individual farmers to work towards the conservation of landraces. **Daniel Putterman**

Road cuts urged to combat warming

London. Britain's Royal Commission on Environmental Pollution last week recommended halving the current UK road-building programme in order to achieve the sharp reduction in carbon dioxide emissions from cars and lorries needed to help stem the rising concentration of greenhouse gases in the atmosphere.

The commission also proposed shifting road policy away from the private car and towards public transport. And it recommended that fuel duty should be increased by another 4 per cent on top of the 5 per cent year-by-year increase in real terms to which the UK government is already committed. That would lead to a doubling of the price of fuel by 2005.

The report has taken two and a half years to complete and contains 110 recommendations and 17 proposed targets to meet the eight objectives identified by the commission. The objectives include increasing the proportion of trips made by environmentally less damaging modes, achieving standards

of air quality that will not damage human health, reducing noise nuisance from transport and reducing carbon dioxide emissions from transport.

Sir John Houghton, the chairman of the commission, is also co-chairman of Working Group 1 of the Intergovernmental Panel of Climate Change (IPCC). The panel has consistently argued that global carbon dioxide emissions must be reduced well below 1990 levels if the concentration of greenhouse gases in the atmosphere is to stabilize at levels that prevent 'dangerous' climate change.

Figures issued by the UK Department of the Environment in late 1992 show that, although total UK carbon dioxide emissions fell by 10 per cent between 1970 and 1990, emissions from transport increased by 65 per cent. They also show that transport accounts for all the projected increase of 39 million tonnes of carbon a year between 1970 and 2020, with two-thirds coming from private cars. **Maggie Verrall**

Ownership and human genome

SIR — A recent leading article (*Nature* 371, 363–364; 1994) addresses ownership of the human genome and weighs private commerce against the public good. The editor acknowledges the apparently unselfish way in which the academic community has published sequence information in public databases, but questions why a profit-making company would publish its sequence data in the same public forum. Certainly it is not merely a question of commercial versus academic motives. The issue is complex and involves an underlying philosophy regarding basic research. After in-depth analysis of this situation, Merck has taken the policy position that certain information relating to the human genome should not be the subject of exclusive ownership. Why should we take this position?

The sequences of expressed genes have been determined with increasing speed and ease. The main commercial use of sequence information has been for construction of protein expression systems for use in drug screening. In this light, Merck has worked to discover and utilize expressed genes as targets for drug discovery. The results of such work has been protected through the patent processes when warranted, and broadly published in both the literature and electronic databases. Physical resources used in this work have generally been made available to both commercial and academic researchers on a non-exclusive basis. This approach, shared by many others, has proved successful for Merck, and has not detracted from our commercial advances.

The set of expressed human gene sequences that is being defined by the EST (expressed sequence tag) approach, pioneered by Matsubara, Sikela and Venter, represents an important and evolving resource for genome research, and provides a potential basis for future biomedical research. EST sequencing may lead to nearly complete gene identification far in advance of the complete sequencing of the human genome. It is this prospect that is most exciting to both commercial and academic concerns. We agree with the editor that "better medicines must come from the identification of genes whose products are involved in metabolic disturbances linked with disease". Yet no current methodology exists that can reliably capitalize on gene sequence information for drug discovery, except for the rare direct use of the expressed protein as a therapeutic. The route from discovered gene to developed drug is long, expensive and has a high failure rate. Biological understanding of the gene and its products in both physiology and pathophysiology is necessary for the validation of a given gene as a potential drug target. In a few

selected cases, the gene or its protein product may have direct therapeutic utility, but for the vast majority of expressed genes the principal value will be as a research tool, one of many, in the drug development process.

Patent protection should encourage the advancement of technology and the creation of saleable products. A key question centres on the determination of the value of intellectual property whether it may or may not be subject to patent protection. Clearly, in the matter of ESTs, the patentability issue remains uncertain, leading several key commercial sequencing concerns to keep their sequences as trade secrets. This fact alone is not uncommon, nor unexpected. However, it is the use of these secret sequences to secure closed collaborations on their biological investigation, and to secure commercial rights to the biological discoveries that result, that created consternation in the academic and commercial research communities.

Against this background Merck has taken the initiative to create a public database of expressed human genes. By providing the necessary resources to some of the most advanced and experienced laboratories, a large-scale, coordinated public EST sequencing effort can be developed. The total set of human genes contains critical information for the understanding of human biology, but we need to develop new methods to extract that knowledge. The complexity of this process will require the intellectual and experimental resources of the world's best biologists for ultimate success. The resulting discoveries should clarify the true commercial potential of these genes.

Since scientists at any research institution, including Merck, can say that "most research occurs elsewhere", we believe that providing broad access to human gene sequences optimizes the possibility that the information will be used to benefit human health. As this extended process creates intellectual property, patent protection can and should be sought by the inventors, and appropriate commercial activities such as drug development will be encouraged. In summary, we believe that the Merck initiative will accelerate the gene discovery process, and that in turn the development of novel therapeutics to benefit human health will also be accelerated. Over the years Merck has sought to make basic research tools broadly available and to work closely with academic scientists, and this will remain our strategy for the future.

Alan R. Williamson

Keith O. Elliston

Merck & Co. Inc.,

PO Box 2000, RY80K,

Rahway, New Jersey 07065-0900, USA

Drug testing

SIR — A recent leading article in *Nature* (371, 90; 1994) comments on drug testing of athletes. It points out that positive results from tests on well-known athletes are invariably made public, and asks why negative results are not also made public so that the prevalence of drug-taking by athletes can be assessed. A drug-testing policy should properly explain how confidentiality of test data will be protected in the event of a confirmed positive test result; and access to testing data, whether the results are positive or negative, is a potentially litigious area¹.

It would be in any case difficult to ascertain the true incidence of drug use by athletes, because many athletes know when to take drugs, at what dose level, and when to stop in order to test negative at competitions². Indeed, some feel that drug testing of athletes can be effective only if carried out on a random, unannounced basis throughout the year^{3–4}. Random testing, however, may raise legal problems. Moreover, most abuse of supposed performance-enhancing drugs may occur in local sports clubs rather than in the competitive arena.

Even in the best of circumstances, the practice and theory of drug testing suffers from potential flaws and limitations¹. Available testing technology for urine samples has distinct limitations; and it is thus quite important to consider the 'sensitivity' of particular screening procedures as well as the 'specificity' of extant confirming procedures. Positive test results should be confirmed by a different test. Your leading article asks why, if there are always two urine samples, they are not analysed independently at separate laboratories. Confirmation of a positive result by a different test in a different laboratory, or the independent analysis of two samples at separate laboratories, would add protection.

Further rigorous study is needed of the technical issues raised by drug-testing programmes for athletes, covering selection of athletes for testing, the timing of drug testing, consent forms, preparing a list of banned substances, and procedures for the handling, collecting and transporting of drug-testing samples¹. The emphasis, however, should be on educating and motivating athletes not to abuse drugs, rather than simply conducting drug testing and then 'punishing' those testing positive for proscribed substances.

L. Uzych

103 Canterbury Drive,

Wallingford,

Pennsylvania 19086, USA

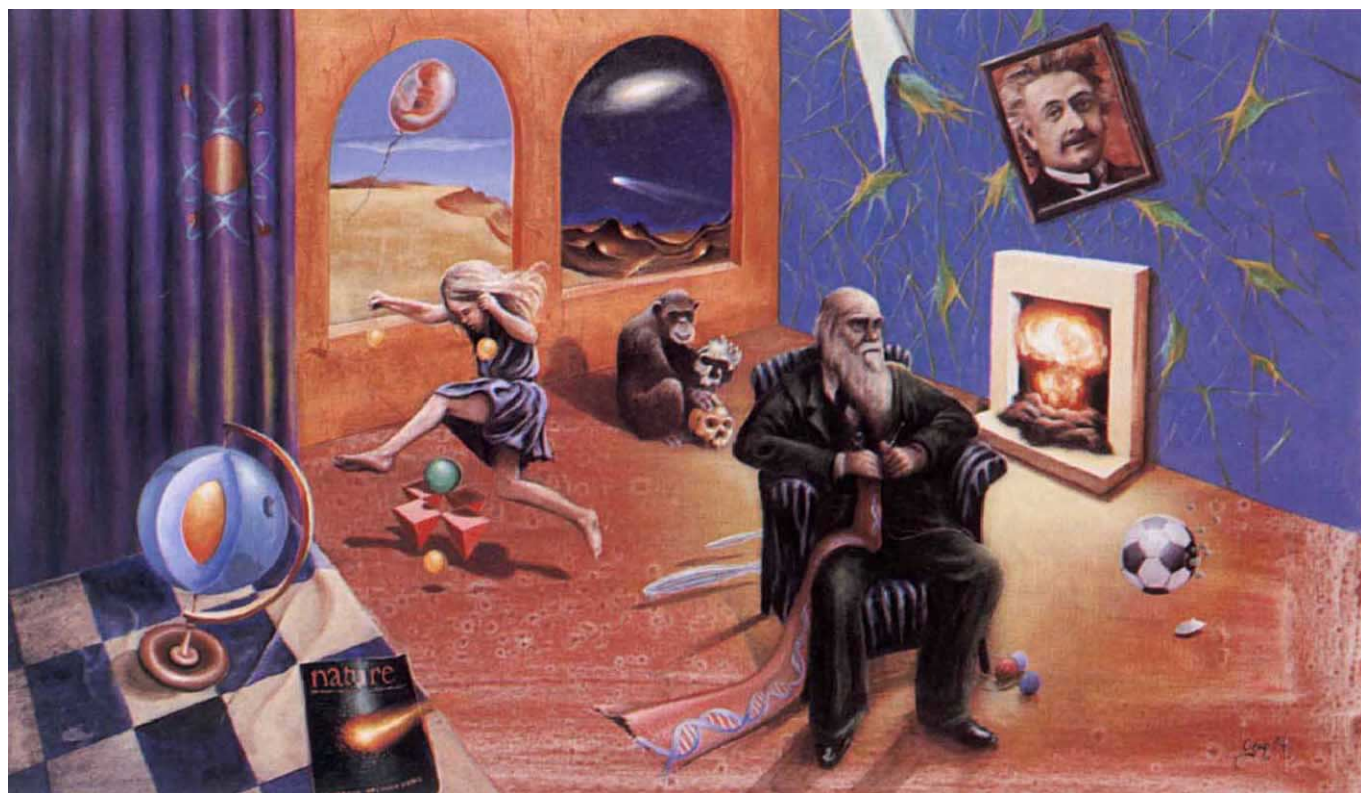
1. Uzych, L. *Br. J. Addiction* **86**, 25–31 (1991).

2. Cowart, V. S. *J. Am. med. Ass.* **260**, 3556–3557 (1988).

3. Cowart, V. S. *Physician and Sportsmedicine* **16**, 165 (1988).

4. Cowart, V. S. *J. Am. med. Ass.* **260**, 3397–3398 (1988).

FRONTIERS OF IGNORANCE



Commission by *Nature* for the supplement celebrating 125 years of science. Greg Griffiths

How to mark 125 years of discovery

THE scientific enterprise has had a marvellous century. Hardly anybody would now deny that. And, indeed, that seemed to be the prospect 125 years ago, when this journal was founded by a group of enthusiasts enlivened by the way in which Darwin's theory of evolution by natural selection had, in the previous decade, transformed general understanding of the natural world. *Nature's* purpose, at the outset, was to tell the world of the "grand achievements" of natural science. But it also took as its cause the improvement of the conditions under which the scientific enterprise laboured, believing that the world at large and British society in particular paid too little attention to the benefits, material and intellectual, that would be wrung from science.

Little is known of the first cohort of readers of this journal, but something may be inferred from what the first editors chose to provide them with. Natural history and archaeology were commonplace. The continued exploration of Africa recurred as a topic. References to documents concerned with public policy in matters such as public education, the administration of universities, important engineering projects and the comparative

economic success of Western European countries were grist for the weekly mill.

There was ample gossip, as when the contenders for important university chairs were listed in advance of the appointment of one of them. Books were reviewed magisterially, as now, but often anonymously. Reports of the proceedings of the learned academies at Berlin, Paris, Philadelphia, St Petersburg and Vienna appeared regularly. And, from the outset, *Nature* was free with its opinions on contemporary issues — the importance of science, its neglect, the meaning of education and the proper regard for "men of science" (quaintly, it must be supposed that women were subsumed therein) forced from their posts, as in Germany in the 1930s.

By inference, the readers of the early issues of *Nature* were a miscellaneous group united by their interest in intellectual affairs rather than in natural science in the strict sense. Many of them may have been the country priests and parsons who had become, in late Victorian Britain, the successors of the monasteries as the repositories of local scholarship and the citadels of regional aspirations to a wider understanding of the world. The painfully

slow, but inexorable, growth of post-secondary education in Britain (with Scotland, as so often, in the vanguard) would have helped the growth of the journal in its early years.

Although engineering had been professionalized early in the nineteenth century, natural scientists were more often the custodians of regional museums, public servants charged with the supervision of technical legislation and interested amateurs than full-time academics at universities. J. J. Thomson, who discovered the electron in 1897, was professor of natural philosophy, not of physics, at Cambridge, for example.

Nature became a scientific journal only by accident. After many decades during which correspondents had written to boast of having experienced especially striking sunsets, remarkable meteor showers or exceptional weather, it must have seemed natural (at the beginning of this century) that correspondents should also describe a theory, a discovery or a scheme for applying new knowledge in some novel way. (That is how *Nature* can now boast of having published the first account of how a television receiver might function a quarter of a century before the event.)

By the 1930s, the tradition of communicating the first news of important discoveries (stimulated nuclear disintegration, the neutron, spontaneous nuclear fission and so on) was well established. We owe to the catholic spirit of the editor for the first 50 years, Sir Norman Lockyer, the opportunity that has since been presented to play a central part in the communication of important discovery. By choice, but of necessity, *Nature* has thus become a professional journal, read mostly by working scientists with a vivid interest in knowing which of their colleagues and competitors will have discovered what this week. It is, of course, pleasing if others find *Nature* as it has become a way of looking over the shoulder of the most productive of the professions.

Lockyer is also the one who gave *Nature* its engagement in events outside Britain. Since his time, those connections have been enormously and deliberately strengthened. The same English-language edition of the journal is now published, in real time, in Britain, Japan and the United States. A weekly edition is also produced in Beijing (with a few weeks' delay). Moscow's *Monthly Nature*, on the other hand, appears just a few days after the end of the month from which its contents are taken. It is a matter for self-congratulation that the British circulation is declining towards 10% of the total as the circulation elsewhere increases. Similarly, the flow of manuscripts for publication is an almost faithful representation of where good science is done. That is as it should be. □

The endless frontier

INDIVIDUALLY and collectively, the discoveries of the past century have transformed our understanding of the world. That proposition is also nowhere denied. It is easily forgotten, for example, that even the concept of the expanding Universe is an invention of the twentieth century. On the other hand, there are few who will overlook the importance of the discovery of the structure of DNA, as recently as 1953, as a means of understanding the way in which all living organisms are sustained in a modified image of their forebears by their genetic material, as well as the patterns of inheritance.

However widely these discoveries are acclaimed, the sheer success of the past century entails dangers, not the least of which is that these few decades will seem to be an isolated and even self-sufficient phenomenon, a period in human history when almost everything there is to learn has been discovered. So much is reflected in the general puzzlement about the future of the world's economy. Now that the microprocessor on a chip of silicon is found in every office, and molecular genetics promises understanding that may lead to the treatment of presently intractable diseases, what further discoveries of importance can there be?

This view of a century complete unto itself is often inappropriately echoed from within the research community itself. Exhilaration at the importance of new developments often tempts even sober practitioners to suggest that this or that field of enquiry is more or less complete, the need for hard work by a small army of researchers excepted. In fields as different as particle physics and molecular genetics, suggestions of this kind repeatedly crop up—and are mistaken, outside the research profession, for triumphalism.

History, too often overlooked in the

practice of science and in the education of young scientists, argues in the other direction. For one thing, the record is replete with occasions when some field of enquiry has appeared to be complete—and when contentment with the explanations on offer has been undermined by the emergence of conflicting data from experiment or observation.

Ptolemy's account of the movement of the planets sufficed for more than a millennium, giving predictions of planetary conjunctions and eclipses well within the accuracy of mediaeval measurement. The tale of its collapse in the two centuries following the death of Copernicus is the heroic beginning of modern science. The end was brought about because of new observations (by Tycho Brahe and by Galileo), but owed much also to the distinctive impulse of modern science to ask "Why?" as well as "How?"

Darwin's theory of evolution was a sea-change of understanding brought about in much the same way. It was not the fact of evolution, but the mechanism, that startled people. "What other could there be?", we now ask.

Then, just a century ago, the world seemed content with Newton's mechanics and even with the concept of the luminiferous aether by which Maxwell's equations had been accommodated within classical mechanics. Who would have guessed that both quantum mechanics and (special) relativity were just a decade away?

These, of course, are the high peaks of discovery. The more common experience is that discoveries amount to refinements of the pre-existing body of understanding. A partly known mechanism may be enriched in texture by the accumulation of further detail, say by the identification of yet another protein molecule with a role

in the regulation of a mammalian gene. Or the emergence of some new data may prove sufficient to justify inferences of a general character. Or a new technique may offer the promise of a harvest of previously inaccessible data.

Most discoveries are of that kind. Only occasionally do they reveal contradictions with what is already known which, by their logic, demand reappraisal of the tenets of some field of enquiry. But the research community is fully alive to the opportunities presented by contradictions to those with the wit to recognize them for what they are. The vigilance induced by the search for contradictions is one of the means by which the research community has won a reputation for meticulous pedantry.

The result is that understanding is never complete, but rather is continually refined. So much is widely understood. Only occasionally, now, does enthusiasm for what seems to be a final peak of understanding induce complacency, blinding even practitioners to the likelihood that it is merely a plateau.

In reality, this feature of the research enterprise is also the reason why science is a cumulative process and why the accumulation of discovery is the collective property of the research community. When Newton wrote to Robert Hooke in 1675 to say, "If I have seen farther than others, it is because I have stood on the shoulders of giants", he was referring to Copernicus, Tycho, Kepler and Galileo, but he was also describing eloquently every scientist's debt to his fellows, competitors and collaborators alike.

It is well understood that the recognition of this truth helps to give the scientific community its distinctive character. Disagreements about the significance of data notwithstanding, the research community is bound together not simply by the comfort of its jargon, but by collegiality of the kind that transcends national boundaries and corporate identities.

That is well reinforced by the necessarily slow pace of important discovery. It is one thing to take as a goal the discovery of when and how human beings or their ancestors developed the facility of language, and quite another to set a timetable for the publication of an explanation. Who can tell whether, and on what timescale, life-forms will be found elsewhere in the Universe? (And who can estimate what the consequences of such a discovery will be for our regard for our place in nature?)

Sometimes, the delays are technical in origin. Without radioastronomy, we would even now not have an understanding of what questions it is worthwhile asking about the Universe. Without the discovery of reverse transcriptase, the human genome projects would be ventures for

some future century (and the biotechnology industry would not exist).

But conceptual difficulties are also an impediment to progress. Newtonian mechanics came into its own only after a succession of giants (Laplace, Lagrange, Euler, Hamilton and Jacobi, for example) had made it usable. Even the publication of the structure of DNA was followed by a dormant decade before molecular biology came to life. How long will it be before the implications of quantum chromodynamics are understood outside a narrow circle of practitioners? When will it be possible to infer from the shape of fossil skeletons which genes with what functions were active in the organisms concerned?

One purpose of what follows is to illustrate the truth of those propositions.

Questions yet unanswered

If the following pages are an attempt to describe the present state of science, they may nevertheless define the areas of our present ignorance, much in the way that a mould in a foundry will define the shape of the object to be cast in it — by exclusion, as it were. But what follows immediately is a more explicit statement of some of the obvious questions that cry out for answers.

Two particular explanations are nevertheless appropriate. First, this survey is exclusively concerned with intellectual, rather than practical, questions, and is predominantly about the physical sciences. That is deliberate. Nobody would pretend that the harvest of useful applications of science in the years ahead will be less important, or less surprising, than they have been in the past century. They too will change our lives. But they will be the products of understanding already gathered. There is a danger, in present circumstances, that comparatively scant attention will be paid to questions bearing more directly on our cultural appreciation of our place in nature.

The attention paid to the physical sciences may seem excessive, but that is also justifiable. There is a sense in which the life sciences need no recognition at this stage in their development. Molecular biology is busily transforming all of them. The pace of development is unparalleled. And there is little doubt that they will earn the enthusiastic support of the academic community and the wider world for decades to come.

The following questions arise, many of them explicitly in what follows.

How did the Universe (really) begin? The importance of this question is not simply that it would be interesting to know the answer, but that the question is a stringent test of our present understanding of

Another is to describe where science is now, towards the end of a marvellous century. Yet another is to list some of the more obvious puzzles still to be solved, and even the more obvious of the contradictions still to be resolved.

The goal is not to undercut the importance of what has been done in the decades past, but to suggest that what we now understand is but a snapshot of accumulating understanding in a process that will continue for the remainder of human history.

The evocative title above is that of the report to President Harry S. Truman by his science adviser Vannevar Bush, which in 1947 laid the foundations for the growth of research in the United States since the Second World War. □

the physical world. Present cosmology, represented by the Big Bang, hangs to a large degree on the present understanding of what matter consists of, but with observational support from astrophysics as well.

Nothing in what follows should be read as dissent from the Big Bang. Such criticism as is justified of this cosmology may rather consist of the assertion that the theory, when coupled with uncertainty about the time-course of the initial event and the supposition of a period of rapid inflation beginning soon afterwards, in one sense advances the epoch from which data might be gathered to about 100,000 years after the initial event. In Popperian terms, it would be excellent if the Big Bang could be made precise enough to falsify.

The question, "Is there physics after the standard model?" is not as particular as it may seem. The standard model is a theory of the interaction between elementary particles that has been developed in the past quarter of a century, which appears to account for all the phenomena so far observed at particle accelerators and in energetic galaxies and stellar objects, and also for the elementary particles so far recognized.

One serious difficulty, however, is that many of the features of the real world have to be included in the standard model 'by hand', as *ad hoc* assumptions about what the world is like. It may seem churlish to describe the standard model, which is one of the great achievements of the past quarter of a century, as 'incomplete', but particle physicists themselves have set exacting standards.

The incompleteness does not lie in the fact that at least two particles of matter have not yet been found, but in that the predictions of the standard model do not follow naturally from a more general

statement of what space and time are like. That may seem a tall order, but particle physicists have set for themselves the goal of bringing gravity into the equations. If everything should fall neatly into place if and when that has been done, we shall be confident that we know what matter consists of.

Astronomy has flourished since the Second World War, while the exploration of the Solar System has been unexpectedly enlarged (by the use of spacecraft, among other things). But the objects in the Galaxy present several unanswered questions, chief among which is that of the meaning of the extra-Galactic quasars. Do they represent common stages in the evolution of most galaxies, or are they exceptional events? In each case, what is going on?

For the rest, it is an achievement of the period since the Second World War that the understanding of the thermonuclear energy source of stars like the Sun has been so quickly and accurately used to win an understanding of their evolution, the synthesis of the elements included. But a degree of caution is appropriate. With each substantial improvement of the techniques of astronomy, there have been novel understandings. Is it reasonable to believe that the process has now ceased?

We tend to forget that the past few decades have also seen the development of an understanding of how the Earth's surface has been reshaped over the past 3,000 million years or so. Indeed, for the first time, the processes of mountain building and the occurrence of earthquakes and volcanoes have been made a part of rational discourse. That does not mean that all the details of the Earth's surface are now comprehended, or even that there will be no further surprises, but there is every prospect that the surprises will be only small ones.

The same, unfortunately, is not true of the deep interior of the Earth. But the Earth is the most accessible of all the planets, and the condition of the molten core and the solid core within that should help to constrain the evolution of the solar nebula from which all the planets were originally formed.

Perhaps even less well understood is how life originated on the surface of the Earth. The simplest explanation is that it was the consequence of molecular evolution in the presence of a substantial flux of radiation from the still young Sun. There is nothing as yet to make necessary more exotic explanations, for example the population of the surface of the Earth by living organisms from elsewhere in the Galaxy (panspermia). But the truth is that attempts to reconstruct the first stages of the evolution of life are plainly a long way short of their targets.

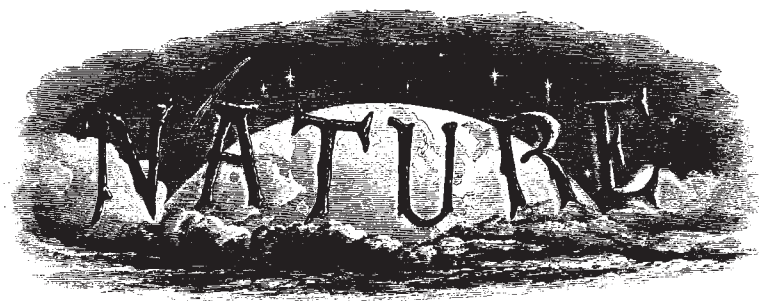
Luckily, this is now less true of the evolution of people on the surface of the

Earth. The australopithecine ancestry of modern human beings is reasonably well charted. But the techniques of molecular genetics may yet make it possible to make inferences about the parallel evolution of people and of the great apes, and even about the genetic basis of the emergence of outstanding human characteristics, that will be of great cultural importance. An equally great prize will be the linking of anthropological and cultural evolution, especially the use of language.

Biology, the booming branch of contemporary science, has questions of a different kind to answer. The most conspicuous of the outstanding problems is that of how the brain works. There is every reason to believe that the outstanding difficulties are to a large extent conceptual. It is not simply that nobody has

yet answered the grand old question, "What is consciousness?", but that even simpler questions such as "What is a memory?" are loosely answered. There are also formidable technical problems, all of them related to the great complexity of the brain, a function of the number of neurons it contains and the number of the connections that each of them makes with others. In this field as in others, biology will have a huge data-retrieval problem on its hands in the years ahead.

None of that implies that the ethical questions already closely linked with biology are unimportant, or that they will melt away of their own accord. The use of genetic diagnosis is already contentious. Luckily, professional biologists seem as fully aware of the difficulties as anybody else. □



A WEEKLY ILLUSTRATED JOURNAL OF SCIENCE

*"To the solid ground
Of Nature trusts the mind which builds for aye."*—WORDSWORTH

THURSDAY, NOVEMBER 4, 1869

NATURE: APHORISMS BY GOETHE

NATURE! We are surrounded and embraced by her: powerless to separate ourselves from her, and powerless to penetrate beyond her.

Without asking, or warning, she snatches us up into her circling dance, and whirls us on until we are tired, and drop from her arms.

She is ever shaping new forms: what is, has never yet been; what has been, comes not again. Everything is new, and yet nought but the old.

We live in her midst and know her not. She is incessantly passing to us, but betrays not her secret. We constantly act upon her, and yet have no power over her.

The one thing she seems to aim at is Individuality; yet she cares nothing for individuals. She is always building up and destroying; but her workshop is inaccessible.

Her life is in her children; but where is the mother? She is the only artist; working-up the most uniform material into utter opposites; arriving, without a trace of effort, at perfection, at the most exact precision, though always veiled under a certain softness.

Each of her works has an essence of its own; each of her phenomena a special characterisation: and yet their diversity is in unity.

She performs a play; we know not whether she sees it herself, and yet she acts for us, the lookers-on.

Incessant life, development, and movement are in her, but she advances not. She changes for ever and ever, and rests not a moment. Quietude is inconceivable to her, and she has laid her curse upon rest. She is firm. Her steps are measured, her exceptions rare, her laws unchangeable.

She has always thought and always thinks; though not as a man, but as Nature. She broods over an

all-comprehending idea, which no searching can find out.

Mankind dwell in her and she in them. With all men she plays a game for love, and rejoices the more they win. With many, her moves are so hidden, that the game is over before they know it.

That which is most unnatural is still Nature; the stupidest philistinism has a touch of her genius. Whoso cannot see her everywhere, sees her nowhere rightly.

She loves herself, and her innumerable eyes and affections are fixed upon herself. She has divided herself that she may be her own delight. She causes an endless succession of new capacities for enjoyment to spring up, that her insatiable sympathy may be assuaged.

She rejoices in illusion. Whoso destroys it in himself and others, him she punishes with the sternest tyranny. Whoso follows her in faith, him she takes as a child to her bosom.

Her children are numberless. To none is she altogether miserly; but she has her favourites, on whom she squanders much, and for whom she makes great sacrifices. Over greatness she spreads her shield.

She tosses her creatures out of nothingness, and tells them not whence they came, nor whither they go. It is their business to run, she knows the road.

Her mechanism has few springs—but they never wear out, are always active and manifold.

The spectacle of Nature is always new, for she is always renewing the spectators. Life is her most exquisite invention; and death is her expert contrivance to get plenty of life.

She wraps man in darkness, and makes him for ever long for light. She creates him dependent upon the earth, dull and heavy; and yet is always shaking him until he attempts to soar above it.

Small issues yet to be decided

The big questions, that stick most easily in the mind, are dealt with elsewhere. The list that follows consists of apparently less important matters, none of which is likely to be decided quickly.

Will the Universe expand indefinitely, or stop?

Will galaxies become more or less numerous?

How long before the radius of the Sun encompasses the Earth?

Is there life elsewhere?

How best to find it?

How big was the australopithecine population?

When did human ancestors learn to speak?

How was the New World first colonized?

Why are human beings predominantly right-handed?

How can external forces make cells divide?

How energy-efficient are living things?

Can human evolution be reconstructed from the genes?

Can a realistic model of a cell be built?

The best cosmology there is

FOR three decades, the Big Bang Universe has been the generally accepted model of how the world we know came into being: between 10 and 20 billion years ago, everything sprang into being out of nothing. The impulse of that event has kept the Universe expanding ever since.

As theories go, the Big Bang has been a great success. It provides a framework within which most observations of, say, external galaxies can be accommodated. It has even accommodated entirely unexpected phenomena, notably the ubiquitous microwave background radiation discovered by Penzias and Wilson (*Astrophysical Journal* **142**, 419–421; 1965).

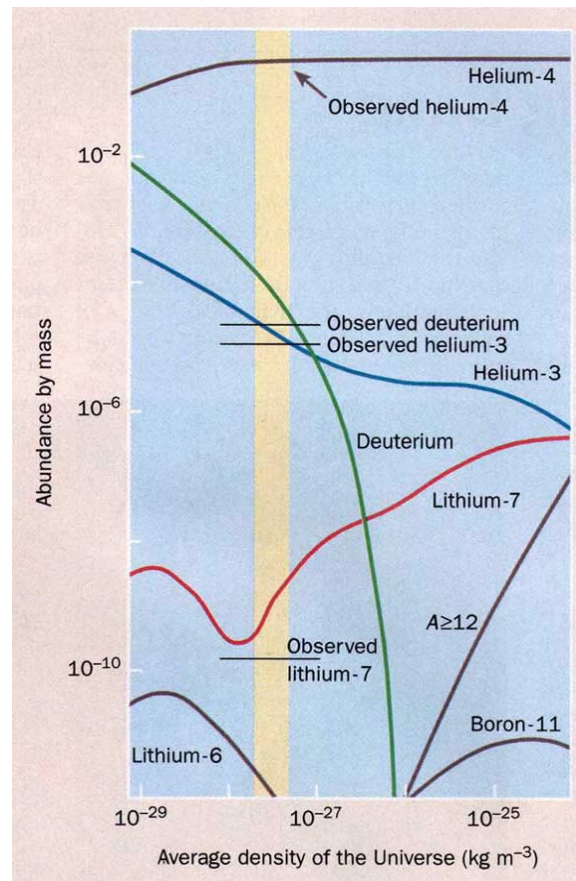
Little is said about the initial event, except that its immediate result was a great concentration of energy, in the form of high-temperature radiation and of particles of matter (consisting of the unstable particles of high-energy physics). Nothing is said, or assumed, about the cause of the event in the sense in which that word was used by David Hume.

Thereafter, the very early evolution of the Universe (occupying just a few seconds) is simply a matter of the succession of intrinsically energetic by intrinsically less energetic particles. Eventually, ordinary atoms (mostly of hydrogen) are formed, perhaps a few seconds after the initial event. At a later stage, perhaps 100,000 years afterwards, the Universe becomes transparent to radiation, which then continues cooling at a rate determined by the expansion of the Universe as a whole.

As a framework for the accommodation of observations, which are mostly of objects such as galaxies formed long after the high-density and high-temperature phase of the Big Bang, this model is not distinguishable from solutions of Einstein's relativity equations that describe an expanding Universe and whose parameters are chosen to match the observed large-scale properties of the Universe, its density for example.

There are two respects in which the observable Big Bang Universe is linked directly with events in its earliest history. The strongest link between the present and the Big Bang is the abundance of deuterium (and other light nuclei) in the Universe as now observed, which is correctly predicted (from known nuclear cross-sections) by the events in the first few sec-

onds of the Big Bang. The present deuterium content of the Universe cannot be the product of nucleosynthesis in the stars, where it is (with hydrogen) a fuel rather than a product, but the Big Bang requires that some of the extant ^3He and ^4He should also be primordial. These arguments imply that the material in the present Universe must at some early stage have been through a high-temperature, high-density phase, and are thus the most direct support for the Big



Predicted abundances of light elements as a function of the average density of matter in the Universe, showing agreement with observations. Figure adapted from *The New Physics*, ed. P. Davies, Cambridge University Press, 1989.

Bang (see figure).

The microwave background radiation, which fills even the corners of the Universe, would psychologically have been more compelling evidence for the Big Bang if it had been predicted before its discovery in 1965. That it was not is something of a surprise, which is nevertheless now irrelevant. The microwave background is a true relic of the time at which the Universe became transparent to radiation.

The uniformity over the sky of the microwave background radiation bears more directly on conditions in the

early Universe than its precise temperature, measured as 2.73 K. Given the inevitable uncertainties of the calculation of the course of events after the Big Bang, the temperature of the radiation must be less critically diagnostic of conditions in the early Universe.

This does not mean that the Big Bang Universe is trouble-free. There is, for example, a problem about the present density of matter in the Universe, estimated from the luminosity and dynamics of visible galaxies to be about 5 per cent of what is called the critical density, that required if gravitational self-attraction is to decelerate the expansion of the Universe asymptotically to zero. (For a Hubble constant at the lower end of the usual estimated range, the critical density is estimated as $5 \times 10^{-27} \text{ kg m}^{-3}$, or roughly one atom of hydrogen per litre).

The difficulty is not that the figure is too far from unity, but that it is uncomfortably close to it. Small departures from equality at the Big Bang would by now have been magnified hugely, so that the Universe would either be conspicuously empty of matter, or conspicuously over-full. What this argument implies is that the Big Bang would have had to have begun with a matter density almost but not quite equal to the critical density to reproduce the present state of affairs. But nobody likes initial conditions so carefully prescribed.

One of the first difficulties to be recognized, in the early 1970s, was the capacity of a Universe expanding as a consequence of an initial impulse to magnify the scale of density fluctuations soon after the initial event in proportion to the expanded scale of the Universe as a whole. The consequence of that would be that the matter in the initial fireball would quickly have clumped, by virtue of its self-gravitation, into galaxies much larger than those now observed. Conversely, the present

large-scale uniformity implies that the density fluctuations in the initial fireball must have been much smaller (perhaps by a factor of 10^{15}) than would be expected for a particle gas in equilibrium.

Both problems have now been resolved, at least formally, by a device called 'inflation', originally due to Alan Guth, for which, since the development of the theories of the weak and strong nuclear interactions, there is now physical justification of a kind. Guth argued, in 1980, that there would be a point in the hierarchical conversion of one kind of

particle into another when something like a phase transition would take place between one regime and the next in the hierarchy, with the release of the equivalent of latent heat sufficient to make the still tiny Universe expand geometrically by a factor of 10^{50} or more in an interval of time no greater than 10^{-32} seconds.

One result of inflation is that it would drive the density of the Universe towards the critical density, whatever the initial conditions. In other words, if there were an inflationary phase in the history of the Universe, the true density of the Universe should now be 20 times greater than it has as been found to be. But 5% is uncomfort-

ably far from 100% and becomes the more so as the search for 'missing mass' remains unsuccessful.

Inflation also irons out the expected fluctuations of the initial high-temperature Big Bang, so that the large-scale uniformity of the Universe becomes explicable, but at the cost of putting an impenetrable observational barrier between the pre-inflation epoch and the present. In other words, it ceases to be possible to tell, from where we are now, the initial conditions for the Big Bang before the onset of inflation. There is a sense in which the Universe began only when inflation ended. □

supposedly entirely devoid of mass. But if all neutrinos had a very small mass, or if some had a slightly larger one, the difficulty of the missing mass would be resolved, given the ubiquity of neutrinos in the Universe. Suggestions⁴ that there may be a neutrino with a mass equivalent to 17 keV were originally welcomed as a means of reaching the critical density, but the case for such a mass (needlessly large for cosmological purposes) is now discounted. In the past 15 years, other material particles have been considered as candidate particles for the missing mass, only to be rejected, usually on the grounds of implausibility. The question of the missing mass remains obdurately open.

Hubble constant. From our present vantage point, the expansion of the Universe is simply described by the Hubble relation between the recessionary velocity of a galaxy (v), perhaps measured by the redshift of a spectral line, and its distance (r). The velocity is a multiple of the distance, or $v = Hr$, where the Hubble constant, H , is measured in units of km s^{-1} per megaparsec. (One parsec, defined as the distance at which the radius of the Earth's orbit about the Sun would subtend one second of arc, is numerically about 3.26

Holes in the Big Bang

THE cosmology that has held sway for the past 30 years commands general support, but that is not to say that it is free from defects. How could it be, a mere three decades after being formulated, and in a field in which the inadequacy of the data is as evident as in cosmology?

What follows is a list of some outstanding problems with the Big Bang picture of the Universe, derived from the published literature. To have compiled it is not to suggest that the Big Bang is an incorrect picture of how the Universe has evolved, but may equally be taken as a recitation of what needs to be done to complete the picture.

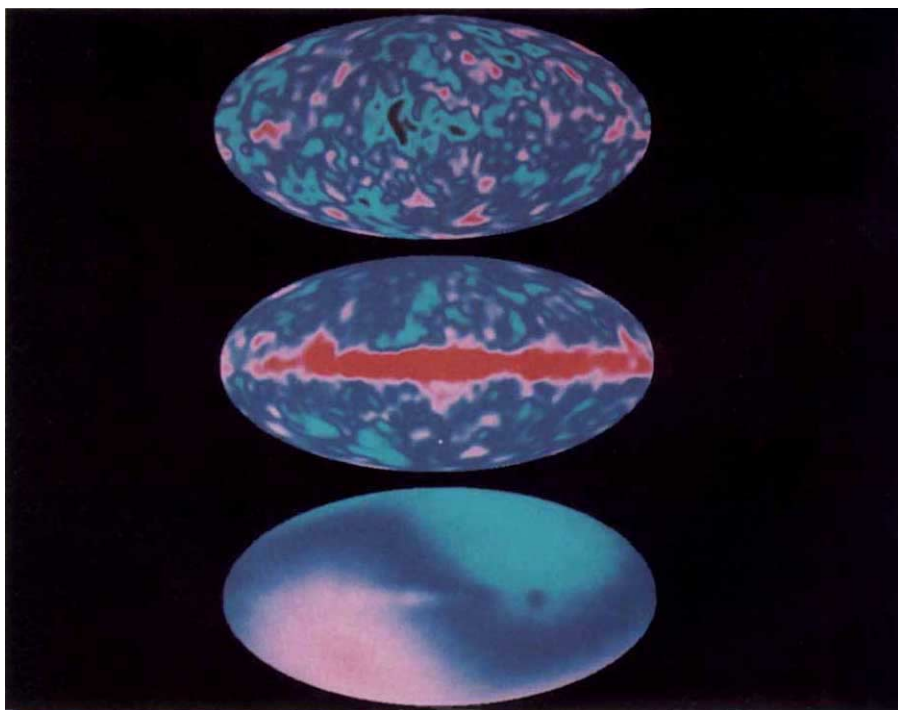
Missing mass. This is a problem only if it is supposed that the density of the Universe is equal to the critical density, ρ_c , say. (Coles and Ellis have recently assembled evidence to suggest that the supposition is untenable¹.) It is conventional (and convenient) to use dimensionless quantities. If the density is ρ and the critical density is ρ_c , the test of whether the density is more or less than the critical value is determined by whether the density parameter $\Omega = \rho/\rho_c$ is greater or less than unity.

The extent of the missing mass is very large. The luminosity of the observable galaxies betokens a mass equivalent to less than 1 per cent of the critical density. On the other hand, accurate measurements of the rotational dynamics of galaxies, and in particular of the rotational velocity of visible material as a function of radius, have shown that the effective masses of galaxies are typically 10 times greater, perhaps enough to account for $0.05\rho_c$.

The unseen mass may plausibly consist of extinct stars, or stars too small to be luminous, in the haloes of galaxies. A recent study of the galaxy NGC5907 has revealed that this may indeed be reality².

On the assumption that $\Omega = 1$, where is the remaining material, and of what does it consist? One possibility is that it is simply primordial gas, mostly hydrogen, left

over from the Big Bang and not yet formed into galaxies. Such material may lie at the heart of galaxy clusters, providing the mutual gravitational attraction that binds them together. But that seems to be unlikely; the dynamical effects of such large amounts of matter would be more pronounced. In any case, calcula-



Full-sky maps, in Galactic coordinates, of variations in 3-K background radiation as measured by the COBE satellite. Top, Variations of 0.12% seen before corrections. Middle and bottom, 0.01% variations that result when first the dipole anisotropy and then the Galactic emissions are subtracted. Along with instrumental noise in the bottom image, the COBE team found variations that suggest density fluctuations in the early Universe.

tions of the formation of ${}^4\text{He}$ in the Big Bang³ place an upper limit on the proportion of primordial matter that can consist of nucleons, which is at least an order of magnitude less than ρ_c . Another possibility is that the missing mass may be embodied in particles such as neutrinos,

light-years; a megaparsec (Mpc) is 10^6 times as much.) To avoid the repetition of the units, it is again convenient (and conventional) to put $h = H/H_0$ where H_0 is taken as $100 \text{ km s}^{-1} \text{ Mpc}^{-1}$, so that h becomes a dimensionless number.

From the outset, in the 1930s, there has

What the ancients believed

COSMOLOGY is historically an episodic field. Over the long period going back to before ancient Greece, the current view of how the Universe is constructed has been drastically changed by the arrival of new techniques and the accumulation of data. Much has changed since some among the ancients likened the stars in the sky to lanterns hung from a spherical and moving ceiling that is itself invisible.

But by the time of Aristarchus (310–230 bc), the Greeks had a clear picture of a heliocentric Solar System, with the distance of the Sun estimated at between 18 and 20 times that of the Moon. Archimedes, in a manuscript of 215 bc, gives an interpretation of Aristarchus in which the Sun is at the centre of a static spherical Universe, with the Earth and the other planets in orbits about it. The fixed stars are fixed.

Erastosthenes, one of the first directors of the great library at Alexandria, provided this cosmology with a yardstick by a measurement of the diameter of the Earth, afterwards much improved by observations by Hipparchus of eclipses, notably the solar eclipse of 129 bc. Hipparchus seems also to have begun a catalogue of the fixed stars, later used by Ptolemy (after the Romans had taken over at Alexandria) in his more extensive compilation of the positions of more than 1,000 stars.

It is far from clear why Ptolemy, apparently¹ following the lead of Hipparchus, had retreated by the second century AD to a view of the world that restored the Earth to centre-stage (and described the motion of the planets by his notoriously complicated system of epicycles), but that was the position of Christendom when, much later, Copernicus reintroduced heliocentric cosmology.

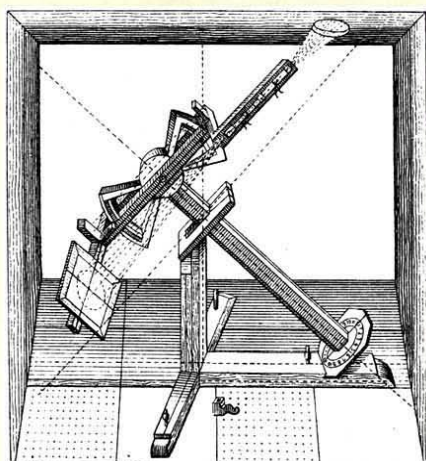
Copernicus's book was published in the year of his death in 1543. The heroic century and a half that followed included Galileo's demonstration of his version of the Equivalence Principle (that gravitational and inertial mass are identical) and culminated with Newton's theory of gravitation in 1687 (published, together with the laws of motion and the differential calculus, in the *Principia*).

The interval was full of drama, in comparison with which arguments among contemporary cosmologists pale into politeness. There was, for example, the burning of Bruno² at the stake in 1600 in conformity with the Inquisition's routine injunction that he should "be dealt with as mercifully as possible and without the shedding of blood" and Galileo's forced abjuration of the heliocentric hypothesis in 1633 followed by his house imprisonment until his death in 1642.

Then came Newton's *Principia*, with a theory of gravity that, for the first time, made possible a rational cosmology, not

to mention the application of newtonian mechanics to the calculation of all physical phenomena, celestial mechanics included. There were also, of course, Newton's quarrels with Hooke (over priority for the theory of gravitation) and with Leibniz (over priority for the calculus).

It is notable that there was nothing in the seventeenth century to suggest that the Universe is other than static. The fixed stars were fixed. But by the end of the eighteenth century, William Herschel was able to measure the proper motion and the parallax of nearby stars, showing that



Telescope of Galileo

the Milky Way has a motion of its own, with individual objects continuing in what Newton had called their state of "rest or motion".

It is surprising, but true, that when William Herschel managed to resolve into stars some of the 5,000 nebulae he had catalogued by 1820, nobody took much notice for the best part of a century. The true scale of the Universe was not appreciated until 1930 or thereabouts, when the 100-inch reflector at Mount Wilson had shown that the Milky Way, "our Galaxy", is mirrored in uncounted others, perhaps 100 million of them, sometimes drawn together into clusters.

The first observations to show that the Universe is not static came earlier, during the First World War, when V.M. Slipher, F. G. Pease and Harlow Shapley, using the new generation of telescopes in the United States, were able to show spectroscopically that the distant galaxies are predominantly moving away from the Sun (or from our Galaxy). Until that point, there would have been no reason for people to abandon the ancient idea that the Universe is fixed for all time, although people such as Jeans were worrying about the longevity of stellar energy sources.

By good luck, that development virtually coincided with the appearance of Einstein's general theory of relativity and was soon followed by de Sitter's demonstra-

tion that one solution of Einstein's equations is consistent with a picture of an expanding Universe.

During the 1930s, the great hunt for signs of life on Mars and other planets, which had occupied some of the best telescope time for half a century, was finally abandoned as people turned to contemplate the vast scale of the Universe revealed by the distance of the galaxies derived from the redshifts measured at Mount Wilson, as well as to astrophysical problems, notably the generation of energy by thermonuclear reactions in stars.

Cosmology came into its own again only after the Second World War, and for three reasons: the 200-inch Hale telescope on Mount Palomar was commissioned in 1950, making the megaparsec the cosmologists' awesome standard unit of distance; the development of radioastronomy made a novel class of galaxies accessible; and Bondi, Gold and Hoyle published their contentious theory of the steady-state Universe in which matter is continuously created.

For the past 30 years, Big Bang cosmology has become generally accepted as the true world view. It may well be a correct one. But the episodic character of the historical record must prompt the question whether an explanation of the expanding Universe, itself exclusively the creation of this century, is likely to be complete.

In the long history of cosmology, we forget how much dead-time there has been. Some of the delays are those needed for novel concepts to become generally accessible. Thus the importance of the *Principia* was not fully appreciated outside Britain until the second edition appeared, in 1713. (The greatest among Newton's contemporaries, Christian Huygens, had already rejected the notion of action at a distance implied by Newton's gravitation.) But then newtonian mechanics had to be made usable by the work of Lagrange, Hamilton, Jacobi and the like. Similarly, it is a phenomenon of recent experience that it has taken the best part of half a century for the general theory of relativity to become a part of physics proper.

On other occasions, cosmology has been kept waiting for the arrival of new techniques. Would it have fallen to Galileo to make the decisive arguments for the heliocentric hypothesis if his predecessors had had access to a telescope with which to see Jupiter's satellites? In our own time, each extension of observation to new regions of the electromagnetic spectrum has coloured (and sometimes changed) our world view. Can we be sure that this process is at an end? □

1. Pannekoek, A. *A History of Astronomy* (Allen & Unwin, London, 1961).

2. Snyder, C. *The World Machine* (Longman, London, 1907).

been great uncertainty about the value of h . That persists. The values measured over the years seem to have clustered around 0.5 and about 1.0. The origins of the uncertainty are well understood: the difficulty of constructing a distance-scale that will span the whole Universe. That requires some means of recognizing in distant galaxies objects whose astrophysical properties are sufficiently well understood within the Galaxy for them to serve as 'standard candles'.

The issue is important because the actual value of h is linked directly to quantities such as the age of the Universe, but in ways that are influenced by the model used to describe the general expansion. At its simplest, h will have decreased since the origin of the Universe because of the way the expansion is decelerated by mutual gravitational attraction (so that the value of Ω will also be relevant).

What has now emerged from measurements of Cepheid variable stars in distant galaxies of the Virgo cluster of galaxies is that the value of h is more like 1.0 than 0.5. Pierce *et al.*⁵ have determined a value of 0.87 ± 0.07 , a result confirmed by measurements with the Hubble Space Telescope⁶. That is inconveniently (but not impossibly) high for the conventional cosmology.

Age of the Universe. Like other properties, the age of the Universe is directly dependent on h (but also on Ω , which determines the deceleration of the expansion). Coles and Ellis use¹ the emerging consensus on such a value of h to conclude that Ω must be considerably less than 1 if the age of the Universe as a whole is not to be considerably less than the ages of the stars of the globular clusters in the Galaxy, which lie in the range 14–16 billion years. The alternative is that if $h = 0.95$ and $\Omega = 1$, there must be a contradiction between the theory of the inflationary Big Bang and such data as the cosmologists have been able to assemble.

Light isotopes. Even the strongest source of empirical support for the Big Bang raises difficulties still to be resolved. Nucleosynthesis in the Big Bang would have begun as the temperature fell to the equivalent of 10 MeV. In present circumstances, free neutrons decay to protons and electrons (plus an antineutrino) with a half-life of just over 10 minutes in a weak interaction transformation involving an energy of 1.293 MeV. As a consequence, at a temperature of 10 MeV, both nucleons are in equilibrium with each other and with electrons and positrons. That allows for the formation of deuterium (^2H) and tritium (^3H) nuclei and, by combinations between the products, of the nuclei of ^4He , ^3He and ^7Li , proportions of which are then frozen out as the temperature of the equilibrium gas falls to the equivalent of 1 MeV or less.

The most recent recalculation of these

reactions³ is based on more accurate measurements of the nuclear reaction rates now available, as well as on recent estimates of primordial abundances (deuterium in the solar wind, helium isotopes in carbonaceous chondrites and ^7Li in population II stars of the galactic halo, for example). The measured abundances can be reproduced, but only if the density of baryons (nucleons) in the Big Bang at 10 MeV is an order of magnitude less than the critical density.

Interestingly, the same calculation yields a limit to the number of types of neutrino allowed; if there were four or more, there would be uncomfortably large amounts of primordial ^4He .

Inflation. The inflationary phase of the Big Bang remains a puzzle, chiefly because its driving mechanism is not unambiguously specified or identifiable in what is known of particle physics. The argument starts from the assumption that the early Universe supported one or more quantum fields of the kind known as 'Higgs fields', which have the property that the state in which the field is identically zero is not that in which the energy is least.

The phase transition responsible for the inflationary phase is therefore a transition from the 'false vacuum' of identically zero fields to the 'true vacuum' in which the numerical values of the field or fields are different from zero. From the latter, energetic particles emerge as the consequences of quantum fluctuations, and set in train the hierarchical sequence of particle transformations that lead, eventually, to the Universe as it is.

The concept is not really as bizarre as it may seem; just such a Higgs field sits astride current theories of particle physics (and the 'Higgs boson' is one of the particles that may yet be found at particle accelerators; see page 23). Making the nature of the inflationary process more specific than it is at present must be an important goal for the future, but one that must depend on a better understanding of outstanding problems in particle physics.

Quantum gravity. A similar (but unrelated) difficulty concerns the treatment of the pre-inflationary phase of the Universe, where it is agreed that the energy density would be so high as to require a quantum treatment of the gravitational field. Much has been learned in the past 20 years of the difficulties of this task, but practical ways of making progress are still beyond the ken of those concerned.

Quasars. The discovery by 1960s radio astronomers of powerful, but apparently point-like, sources of radio energy at great redshift (and thus at an early stage in the evolution of the Universe) is an astrophysical problem with cosmological implications. Several questions arise. Is a quasar a stage in the early evolution of only some galaxies, in which case how are the relics

of that past condition to be recognized in the present Universe? Or is a quasar a stage in the evolution of most galaxies, in which case what is its significance?

Both questions bear directly on the cosmological question, given the importance of galaxy formation as a stumbling block in the proper understanding of the evolution of the Universe without the invocation of inflation.

Alternatives? The formidable problems attending Big Bang cosmology do not imply that the picture must forthwith be abandoned. There are several ways in which the simple version of the Big Bang may be varied, many of which have been aired in the literature. But there seems no easy escape from the contradiction between, on the one hand, the arguments in favour of $\Omega = 1$ and, on the other, the low value of the baryon density required by nucleosynthesis and the disappointment, at least so far, of the search for non-baryonic dark matter.

It is true that Hoyle, Burbidge and Narlikar have recently⁷ argued for what they call the "quasi-steady-state cosmology", which entails episodes of matter creation that tend to be concentrated at 40-billion-year intervals in the history of a Universe that is infinite in space and time.

That has the advantage over the earlier steady-state theory of Bondi, Gold and Hoyle in providing a means by which nucleosynthesis could proceed. It can also accommodate difficulties arising from objects in the present Universe that are older than the age of the Big Bang itself, which may be objects left over from an earlier phase of matter creation. The difficulty is that is hardly more specific about the quantum field supposed to be responsible for the appearance of matter than are the theories of the inflationary phase about the Higgs fields they require. Does the future of cosmology rest with the particle physicists? Other alternatives hang on the notion that string-like structures in the early Universe may have evolved into large-scale structure; "cosmic string" is the name.

Meanwhile, it is relevant that continuing uncertainty about the value of Ω means that we do not know whether the Universe will continue to expand indefinitely (if $\Omega < 1$), or will decelerate in its expansion to the point at which it collapses again under its own weight (if $\Omega > 1$). □

1. Coles, P. & Ellis, G. *Nature* **370**, 609–615 (1994).
2. Sackett, P. D., Morrison, H. L., Harding, P. & Boroson, T. A. *Nature* **370**, 441–443 (1994).
3. Walker, T. P., Steigman, G., Schramm, D. N., Olive, K. A. & Kang, H.-S. *Astrophys. J.* **376**, 51–69 (1991).
4. Simpson, J. J. *Phys. Rev. Lett.* **54**, 1891–1893 (1985).
5. Pierce, M. J., Welch, D. L., McClure, R. D., van den Bergh, S. & Racine, R. *Nature* **371**, 385–389 (1994).
6. Freedman, W. L. *et al.* *Nature* **371**, 757–762 (1994).
7. Arp, H. C. *et al.* *Nature* **357**, 287–288 (1992).

One discovery inside another

ARISTOTLE shared with Plato the view that different kinds of matter are composed of earth, air, fire and water in appropriate proportions. Understandably, the ancients concentrated on the qualities of matter. A brief chronology of succeeding events will clarify where we are now.

■ In the four decades between Lavoisier and Dalton, the atomic theory was established. The size of an atom (of the order of 10^{-10} m in diameter) and the periodic table arrived only in the last third of the nineteenth century.

■ J. J. Thomson announced the discovery of the electron in 1897, Rutherford established the nuclear atom in the succeeding decade and Bohr published his semi-classical model of the quantum atom in 1913. The transmutation of elements by natural radioactivity was a proof that nuclei are not immutable.

■ Dirac, having (in 1928) made a relativistic version of Schrödinger's equation for an electron in an electromagnetic field, famously predicted the existence of a positively charged electron; the discovery of the positron followed (Anderson, 1931). Each particle except the photon (which is its own antiparticle) has such a *Doppelgänger*. The hunt for unstable particles in cosmic rays was thereby intensified, while Yukawa (1937) proposed that forces between nucleons are mediated by particles called mesons (then known to the cosmic-ray community as 'mesotrons') which, although massive, would otherwise function much as photons mediate electromagnetic forces.

■ W. Heisenberg, in 1932, proposed that neutrons and protons are but two almost degenerate states (in the quantum sense) of a single entity, called the nucleon. This step was conceptually important because (a) it introduced a new quantum number, called isotopic spin (otherwise isospin) into the description of nuclear matter which, after nearly 30 years, (b) led directly to the notion of the quark.

■ C. F. Powell and his colleagues¹ used the ionized tracks of cosmic rays in photographic plates to demonstrate, in 1947,

that the candidate mesons in cosmic rays are of two kinds, the pion and the muon. (The former is the supposed mediator of the internucleon force; the latter turns out to be a heavier version of the electron.) In

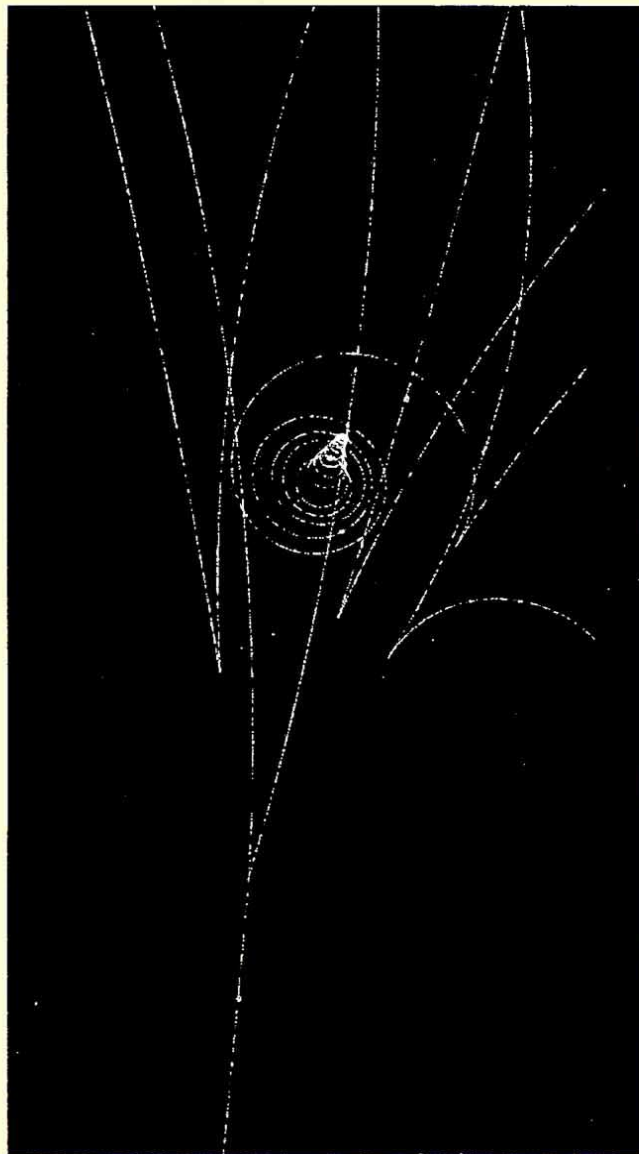
Brookhaven National Laboratory confirmed the reality of a particle called Ω^- , whose existence had been conjectured only.

■ Meanwhile, in the 1950s, theoreticians (Feynman, Schwinger and Tomonaga among others) were developing a quantum version of electrodynamics that satisfied the requirement that it should also be relativistic. Quantum electrodynamics (QED) was a conceptually crucial step. There are two interacting entities: an electron field and an electromagnetic field, represented by its four-dimensional potential. Quantization of the latter ensures that electromagnetic energy exists in the form of photons, with energy $h\nu$, where h is Planck's constant and ν the frequency. But both the electron field (represented by a wavefunction) and the electromagnetic potential contain elements of ambiguity; the former, for example, can be multiplied by a complex function of unit modulus without changing the physical interpretation that the squared modulus of the wavefunction represents electron density. The potential is similarly ambiguous. So the numerical values of the two interacting fields (or, more correctly, the phases) can be redefined at all points in space and time. That is the gauge principle, central to all later theories of particle fields.

■ The prediction of the existence of the Ω^- had been based on a classification of newly created unstable particles made independently by Gell-Mann and Ne'man. Part of their purpose was to account for the multiplicity of the particles already found; high-energy physics needed a periodic table, which in retrospect is a way of classifying the known atoms by means of quantum numbers of which Mendelev was necessarily ignorant. (The first row of the table is filled by the elements H

and He because the Pauli exclusion principle prevents more than two electrons from occupying the lowest electron state.)

That breakneck series of developments is the origin of particle physics as it had become by 1960 or thereabouts. □



A positive kaon (K^+) comes in from the bottom of this bubble-chamber picture, hitting a proton and producing a shower of subatomic particles. The kaon was the first of a whole family of 'strange' particles identified in cosmic rays. (Reproduced from *The Particle Explosion* by F. Close, M. Marten & C. Sutton, Oxford University Press, 1987.)

the same year, the particles called 'strange' were discovered² in cosmic rays, heralding the coming of the quark.

■ The commissioning of particle accelerators producing protons with energy of the order of 10^9 electron volts (eV) made possible the artificial production of antiprotons (with negative, not positive, electric charge). More significantly, in 1964 an experiment at the accelerator at the

1. Lattes, C.M.G., Occhialini, G.P.S. & Powell, C.F. *Nature* **166**, 453–456 (1947).

2. Rochester, G.D. & Butler, C.C. *Nature* **166**, 855–857 (1957).

Building-blocks of matter

THE outcome of the formative period of particle physics is familiar. Matter is made of quarks and leptons.

Hadrons. Combinations of quarks make hadrons, which are either baryons (which, like electrons, have half-integer spin) or mesons (whose spin is either zero or integral). There are six quarks, with the whimsical names up, down, strange, charm, bottom and top, for each of which there is also an antiparticle with an electric charge which is opposite in sign.

Quarks resemble electrons in having half-integer spin. Why they differ from all other particles of matter in carrying charges that are fractions of the electronic charge is not understood, but the up quark has a charge of $+2/3$ units, explaining why two ups and a down (with charge $-1/3$) make a proton and also why two downs and an up make a neutron.

Mesons are simpler. For example, the pion (π) consists of pairs of up and down quarks (called u and d in what follows), but not straightforwardly; such a combination would have non-integral charge. Thus the π^- is a combination of an anti-u (written $\bar{u}$) with a d, while the π^+ is the combination of a u and a $\bar{d}$, giving charges of -1 and $+1$ respectively. The π^0 , by contrast, is a mixed state of $u\bar{u}$ and $d\bar{d}$ (which is one reason why its decay is 10^8 times more rapid than those of its two partners); the predominant decay products are two energetic photons, recognizable by the electron-positron pairs they eventually produce. For all three, the total spin is zero.

The problem with this simple view (as it

formed in the early 1960s) is that there are too few quantum numbers to account for the variety of the unstable particles that had then been created by the accelerators. (The Pauli principle is an efficient way of exhausting the capacity of classifiers.) How to resolve that difficulty? With the

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Computer reproduction of a hadronic decay of a Z^0 particle as seen in the OPAL detector at CERN's Large Electron-Positron collider, LEP.

discovery of the strange particles (in 1947), 'strangeness' was added as an additional quantum number. Until the discovery of the charmed quark in 1975, three quarks seemed sufficient to account for the hadrons then known.

Leptons. Electronic matter is similar, but different. Just as there are three pairs of quarks ((u and d), (s and c), (b and t)), so there are three 'generations' (as they are known in the trade) of leptons — e, μ and

τ — and the corresponding neutrinos, ν_e , ν_μ and ν_τ , all of them electrically neutral and supposedly without mass. These particles interact with others of the same generation and also with quarks by means of the 'weak' nuclear force.

While nobody knows why nature accommodates both the electron (e) and its two more massive congeners μ and τ , the theorists are lucky that these particles transform among themselves. Thus the muon is spontaneously transformed into an electron by $\mu^\pm \rightarrow e^\pm + \nu_e + \bar{\nu}_\mu$. That is the weak force in its pure and simple manifestation.

After quantum electrodynamics, the outstanding triumph of theoretical particle physics is the 'electroweak' theory of Glashow, Salam and Weinberg in the late 1960s. The interacting fields are those of leptons, quarks and of the particles analogous to photons that mediate their interaction.

The surprise is that the mediating field is that of the particles called intermediate vector bosons, or Z^0 and $W^\pm$, predicted before their discovery at CERN in 1983, whose masses are very large. The large mass of these particles is one reason why, at low energy, the 'weak' force is indeed weak. The decay of the $\mu^\pm$, for example, is not a direct conversion, but involves the intermediate (and 'virtual') production of $W^\pm$ (and of ν_μ) followed by its conversion into $e^\pm$ and the electron neutrino. The intervention of particles such as $W^\pm$, whose energy is much greater than that liberated by the conversion of $\mu^\pm$ into $e^\pm$, is allowed in the calculations because of the simple rule of quantum mechanics that any state that is not a pure state (or eigenstate) must be, in principle, a combination

CERN Photo

The ghosts in the machine?

THE great neutrino puzzle has dragged on for a quarter of a century, and is not yet in sight of resolution. The presenting symptom is the deficit of neutrinos emitted by the Sun, first measured by R. Davies Jr by means of an experiment first planned 30 years ago¹ and designed to provide direct evidence of the thermonuclear reactions in and near the Sun's core. It consists of a tank filled with 600 tonnes of the cleaning fluid tetrachloroethylene installed beneath 1,500 metres of overburden in the Homestake mine at the town of Lead in South Dakota.

The relevance of the cleaning fluid is its content of the isotope ^{37}Cl , with which electron neutrinos should infrequently

interact to yield radioactive ^{37}Ar by the reaction $\nu_e + ^{37}\text{Cl} \rightarrow ^{37}\text{Ar} + e^-$. Atoms of ^{37}Ar are collected every two months or so, when their concentration represents a balance between production and decay. The conversion of ^{37}Cl to ^{37}Ar , which is an endothermic process (as chemists would say), can detect neutrinos only when their energy exceeds 0.8 MeV.

Over a quarter of a century, the measured flux has been consistently only a third of that expected. Expectations are almost independent of the fine details of the solar model, but are almost entirely determined by the rate of conversion of hydrogen and deuterium into heavier elements and thus by the

total energy output of the Sun.

This measurement has since been repeated with other pieces of equipment. Of those so far mounted, the Japanese experiment called Kamiokande (based on a tank of 6,000 tonnes of pure water in a deep mine) is the most interesting. It relies on the expected scattering of neutrinos by electrons, which are themselves detectable by a pulse of Cherenkov radiation if their speed is greater than that of light in water.

The experiment has the advantage that it is possible to win directional information about the neutrinos responsible for knocking electrons out of atoms (confirming that those detected do indeed come from the direction of the Sun) as well as to estimate their energy. Because the scattered electrons can be detected only when

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

A fish-eye view of the inside of the Kamiokande detector, looking upwards.

their speed exceeds that of light in water, the Kamiokande experiment detects solar neutrinos whose energy exceeds 8 MeV. The measured detection rate is consistently half of that expected, although the discrepancy decreases with increasing energy.

The measurements so far undertaken all suffer from the disadvantage that they do not detect the low-energy neutrinos expected to be the most prolifically produced by the Sun, those arising from the simple fusion of pairs of protons by the reaction $p + p \rightarrow d + e^+ + \nu_e$ (where d is a deuteron). That deficiency will be remedied by other experiments now being planned, but there appears to be no escape from the conclusion already reached that the solar flux of neutrinos is too small by a substantial fraction.

How serious a setback for the standard model of particle physics would that conclusion be? Probably not very serious, but it would add a degree of empiricism to what claims to be a self-standing theory.

As things stand, it is far from clear why there are three separate kinds of electron — not only the ordinary $e^\pm$, but the $\mu^\pm$ and the $\tau^\pm$, together with the corresponding neutrinos, ν_e , ν_μ and ν_τ and their antiparticles. (The μ^+ is the antiparticle of the μ^- just as the positron is the antiparticle of the electron.) The neutrinos have no electric charge and, conventionally, are supposed also not to have mass. Moreover, as representatives of separate lepton generations, it is supposed that they do not transform directly into one another.

Yet the only plausible explanation of the neutrino deficit so far advanced is that there is a process whereby the three states of neutrinos 'oscillate' among themselves in such a way that the flux of solar neutrinos

reaching the Earth is more or less an equal mixture of all three. Because the μ and τ neutrinos interact less readily than the ν_e with other material particles, the result is that there is a deficit in the numbers recorded by the experiments. On some versions of this argument, one or other (and perhaps both) of the neutrinos should have a very small mass, perhaps a fraction of an electron-volt, which may be

The left-handed Universe

RECORDED speculation about God's professional qualifications have Him as a mathematician (Einstein) and a theoretical physicist (Hoyle), but one particle-physicist wag is reported to have said that He must, in any case, be left-handed. The reference is to the geometrical asymmetry found among the phenomena governed by the weak interaction in the 1950s, and to the circumstance that the electron neutrino is inherently left-handed.

The problem is that known as 'CP invariance' following T. D. Lee and C. N. Yang, who argued in 1956 that the two alternative modes of decay of the charged kaon could be explained only by discarding one of the rules of simple mechanics — that the laws of physics are not changed by an inversion through the origin of the coordinates.

The 'P' in 'CP' stands for 'parity'. In this connection, it refers to the behaviour of the wavefunction describing some system under an inversion of the coordinates (when each is replaced by its negative). The concept crops up most commonly in spectroscopy, where electronic states are labelled as 'even' or 'odd' according to

relevant to the question of the 'missing mass' required to 'close' the Universe.

Minor though those changes may appear, they raise difficulties for the 'electroweak' theory, without question one of the great conceptual triumphs of the past three decades, in which it is supposed that leptons do not interact directly with other particles, but only through the agency of the intermediate vector bosons, $W^\pm$ and Z^0 .

Thus the muon is spontaneously transformed into an electron by the reaction $\mu^\pm \rightarrow e^\pm + \nu_e + \bar{\nu}_\mu$ and not, as might be expected, by $\mu^\pm \rightarrow e^\pm + \gamma$, where γ represents a photon. The explanation is that the transformation must proceed through the intermediary of a 'virtual' $W^\pm$, whose relatively huge mass ensures that the transformation will be infrequent and the lifetime of the $\mu^\pm$, at least by the standards of unstable nuclear matter, will be long.

In the end, the problem of the neutrino deficit will be settled by experiment. Several of the newly designed neutrino detectors are intended to provide direct tests of the oscillation phenomenon. The consequences for theory are less obvious. Grand Unified Theories could accommodate neutrino oscillation, but also predict entities such as magnetic monopoles and phenomena such as proton decay that have not yet been observed. □

1. Davies, R. Jr *Phys. Rev. Lett.* **12**, 303–305 (1964).

whether the wavefunction remains unchanged or acquires a minus sign on inversion. Indeed, parity functions as a kind of geometrical rather than dynamical quantum number.

The trouble with the $K^\pm$ is that it decays in two apparently inconsistent ways, either into two $\pi^\pm$ s and a π of opposite sign, or into a $\pi^\pm$ and a π^0 . That looks wrong. Kaons differ from pions (π^+ , π^0 and π^-) in having a strange quark in their constitution, but otherwise they have negative intrinsic parity like the pions. Because the straightforward dynamical decay of a particle cannot complicate the conservation of parity by creating angular momentum, it follows that parity is not always conserved by the weak force.

That the weak force does not respect the conservation of parity has since been amply confirmed, both by particle experiments and, interestingly, by observations of the direction in which electrons are emitted in the radioactive decay of nuclei whose nuclear spin is predominantly oriented in one direction by an external magnetic field. The first experiment, with ^{60}Co nuclei, revealed the

asymmetry predicted by Lee and Yang.

Since then, it has emerged that the neutrinos resulting from the weak interaction are invariably left-handed, in the sense that the direction of their intrinsic spin points backwards along the direction of their motion. (By the same test, anti-neutrinos are invariably right-handed.) Linked with that asymmetry is the polarization of the electrons from β -decay, which are also predominantly left-handed, at least if their velocity is sufficiently great. Those properties of the weak interaction have now been amply confirmed.

This state of affairs is not the assault on first principles that it may appear to be. Although other phenomena appear to be invariant under parity reversal, there is no *a priori* reason why the weak force should not be a left-handed force. But this empirical discovery of the property of the world we live in must be included 'by hand' in theories describing the weak interaction.

From the outset, it was recognized that the chirality of the weak interaction would force another compensating asymmetry, which turns out to be that of charge conjugation (the 'C' in 'CP'), most simply described as that of replacing every particle by its antiparticle in some physical system. C-invariance would require that the new system should behave exactly as does

the old. In reality, the weak interaction does not respect this principle either, but it does seem that the effects of the weak force under the consecutive application of C and P (in either order) leaves physical systems unchanged. CP-invariance is a natural law.

Or almost so. Evidence to the contrary has emerged in the past six years from particle experiments at CERN and elsewhere. The test-bed for these enquiries is the interactions of the K^- and K^0 mesons. The exact degree of CP violation (if any) in the real world remains an open question, although the theorists appear to have no great difficulty in accommodating any degree of CP violation in the electroweak theory by adjusting the parameters of the Higgs fields.

The immediate need is to pin down the degree to which CP violation is a reality. That argues for still more elegant experiments at the particle accelerators. But, for other reasons, it seems inevitable that the years ahead will see a further increase of attention being paid to the direct detection of neutrinos. It may prove not to be entirely without interest that a degree of CP violation would, in principle, allow the detection of distant parts of the Universe that consist of antimatter rather than ordinary matter. $\square$

colliding elements were billiard balls. There followed a decade intended to explore the internal structure of the nucleon, believed to contain what were called 'partons'.

The construction of a theory of the structure of nuclear matter was nevertheless guided by the pre-existing electroweak theory, and virtually coincided with the discovery of the charmed quark at the Brookhaven National Laboratory in 1975.

Again it was a matter of classifying quarks, and of deciding what attribute of a quark serves as the equivalent of the electric charge that determines the electromagnetic interaction between charged particles, as well as of identifying the particles that mediate the force.

The outcome of that enquiry seemed at the time to be particle theorists' equivalent of the conjuror pulling a rabbit from a hat. The particles that mediate the force are gluons, eight supposedly massless entities. The strong interaction's equivalent of electric charge is similarly a three-valued attribute called colour. (Conventionally, the colours are called 'red', 'green' and 'blue', but 'black', 'white' and 'grey' would serve as well, as would the labels '1', '2' and '3'.) All six particles of the three pairs (or 'flavours') of quarks can thus carry any one of the three colours. As with electric charge, the rule is that like colours repel each other and unlike colours attract.

So should there not be nine (3×3) distinct versions of the π^+ meson, one for each assignment of one of three labels to each of two quarks? Or twenty-seven ($3 \times 3 \times 3$) versions of each nucleon? That, of course, is not how it turns out. Having invented colour as a novel quantum number, particle physicists have been compelled to restrict the number of entities that can occur to those that are unchanged, except for a change of sign, when the colour labels are exchanged between any pair of particles. This is another way of saying that pairs of quarks in an antisymmetric state attract one another.

What this means for the version of the π^+ found in the real world is that it is a mixture, in the sense of quantum mechanics, of the three states in which both the d and the u quarks have the same colour labels. Similarly, the neutron, a combination of two ds and a u, is a combination of the nine states obtained by permuting the three colour labels.

All this seems a little like pulling a rabbit from a hat only to banish it again. It is therefore relevant that the Ω^- particle predicted by Gell-Mann is a curious combination of three strange quarks whose total spin is 3 half-units, suggesting not merely two but three identical particles in the same state and thus an outrageous violation of Pauli's principle. The accepted explanation is that the wavefunction of the state is the unique antisymmetric com-

The glue within the nucleus

UNTIL the discovery of the strange particles in 1947, there was every reason to believe that nucleons have no internal structure. Electrons are (still) believed structureless; why not the proton and neutron also? But the strange particles, which are produced (in pairs) by the collision of nuclear matter — π mesons and protons,

for example — inevitably raised the question of whether there might be more fundamental elements within them.

More direct evidence of internal nucleon structure came from the recognition, at particle accelerators, that high-energy scattering of electrons and mesons by protons could not be described as if the

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Mock-up of the LEP tunnel with the Large Hadron Collider (LHC), still in the planning stages. Following the demise of the Superconducting Super Collider, the LHC is the best hope of finding the Higgs boson (see page 23).

CERN Photo

bination of the ten states obtained by permuting the three colour labels among three otherwise identical strange quarks.

The rule yields just the colour combinations that can naturally occur, no more and no fewer. But there are also more direct tests of the idea. Among other things, the magnetic moment of the neutron is a direct consequence of the spin of the u quark, given that the Pauli principle requires that the spins of the two accompanying d quarks must be oppositely aligned. Moreover, there are many particle collisions in which the likelihood of a particular outcome is a function of the number of colour states of the underlying quarks, so that the scale of the quantum number can be measured directly.

Successful though quantum chromodynamics (QCD) may be in classifying hadrons, it is less successful as a means of accurate computation than the electroweak theory, and for two inescapable reasons. First, the interaction it deals with is the 'strong' interaction, which leads directly to the consequence that the first approximation to the solution of equations (by perturbation methods) may not be not a good approximation. (The electroweak theory has the benefit that the successive terms in approximate calculations are proportional to successive powers of the dimensionless fine-structure constant $2\pi e^2/hc$, numerically approximately $1/137$.) Second, because the particles that mediate the force, the gluons, carry colour charge, they interact with each other, giving the theory an awkward nonlinearity.

The computational difficulties have led to a way of calculating QCD processes by means of lattice models, in which the fields of quarks and gluons are represented by their (variable) values at the vertices of a manageably small lattice. (The technique was developed by K. G. Wilson, who was awarded a Nobel prize for his work in 1982.) Perhaps the remarkable feature of this technique is that it yields good approximations so often.

Whether the theory is philosophically compelling is another matter. As with the electroweak theory, which would allow the existence of the unphysical right-handed neutrino, QCD could accommodate a greater range of particles than are known to exist. The origin of the restrictions found in nature is empirical, not a natural feature of the theory. Moreover, the theory will not be tested properly until the top quark is definitively found; hints of the discovery earlier this year from Fermilab have yet to be confirmed. The latest estimates, inferred from the production of more than a million Z^0 and $W^\pm$ particles at CERN, suggest that the mass of the top quark is within 10 per cent of 172 GeV. □

Vital yardstick hidden still

THE Higgs boson, not yet discovered, is arguably the most important of all particles in contemporary particle physics. To the extent that it is an indispensable part of the electroweak theory of the electromagnetic and the weak nuclear forces, the theory will not be properly tested until the particle has been found. That is especially important because the Higgs particle is a strictly theoretical construct.

One practical difficulty is that there is no unambiguous estimate of the mass of the Higgs particle, although the failure to find it so far suggests that the mass must exceed that of, say, the W^- , now copiously produced at particle accelerators of sufficient energy.

The role of the Higgs particle is abstract: to create the asymmetry between left- and right-handed particles that is the hallmark of the electroweak theory. How can the mere existence of a particle (or, more strictly, of the field of which the particle is a manifestation) accomplish that? It can, if the state in which the field is everywhere zero is not the state of lowest energy.

That sounds wrong, but is not impossible. Particle physics is ready with homely examples. Thus a ferromagnetic material below its Curie temperature will usually consist of domains in which atomic nuclei are ferromagnetically ordered even when the overall magnetization is zero. In the process, both the translational and rotational symmetry of the solid are lost. The symmetry is said to be "spontaneously broken".

In the electroweak interaction, the non-zero Higgs field is the analogue of the

domain-structured texture of a ferromagnetic material below its Curie temperature. The asymmetry is the origin of the left-handedness of the neutrino. But the Higgs field also interacts both with the massless photon and with the intermediate vector bosons, W^- and Z^0 . One effect of that interaction is to give these mediating particles their characteristic (and very different) masses.

Plausible though it may be, the concept of the Higgs field has startling implications for what is meant by the physical vacuum, and in particular for the notion that the vacuum is literally empty. In the immediate neighbourhood of a lepton, it appears that there will be a non-zero Higgs field whose presence determines the magnitude of its interactions with other particles.

The idea that the physical vacuum may be the source of physical particles is not all that novel, of course. A sufficiently energetic photon will generate electron-positron pairs, for example, at least if there is a charged object such as a nucleus in the neighbourhood. A more luxuriant illustration of how the vacuum may be a source of matter is Hawking's account of the production of pairs of particles in the strong gravitational field of a gravitationally collapsed object, a black hole.

The novelty of the Higgs mechanism is that its *raison d'être* is so abstract. To demonstrate the reality of its existence would not merely provide the missing capstone for the electroweak theory, but also a more tangible starting point than there now is for the manipulation of the theory. □

Putting four forces together

UNIFICATION has been in the air since the 1930s, when Einstein and Heisenberg separately advertised ambitions of this kind. Einstein, encouraged by the success of his theory of general relativity, was looking for similar ways of describing the quantum fields of particle physics. Heisenberg's goal, at the outset more limited, was to use similar language to describe nucleons (protons and neutrons) and fermions (electrons and their congeners).

A much more ambitious goal is now in the air: the unification of the four fundamental interactions in which particles can be engaged. Part of the enthusiasm stems from the success of the electroweak theory of Glashow, Salam and Weinberg, which uses a single framework to describe the behaviour of two interactions — the electromagnetic and the weak nuclear forces.

There has indeed been further progress

in that direction, notably with the gauge theory of quantum chromodynamics, which is the theory of the strong nuclear interaction that binds quarks together in hadrons. (The strong interaction acts exclusively between quarks, and thus differs from the weak interaction, which can 'see' quarks as well as leptons.) So why not put them both together, for a start?

That is not just an idle ambition, but consonant with the expectation that, at high energy, a common description will be possible. After all, the 'strong' and the 'weak' nuclear interactions stand out from the electromagnetic force because of their short range. But the range of the weak force, for example, is a simple consequence of the large mass (~ 80 GeV) of the $W^\pm$ bosons. In interactions in which the energy exchanged is much greater, the huge difference of mass between the bos-

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

The NUSEX proton decay detector, in the Gran Sasso tunnel beneath Mont Blanc. Centimetre-thick iron plates sandwich rows of 'streamer tubes', which produce a pulse each time a charged particle passes through. (Reproduced from *The Particle Explosion*, by F. Close, M. Marten & C. Sutton, Oxford University Press, 1987.)

ons and the photon should be irrelevant.

That is the goal of Grand Unified Theories, in the whimsical tradition of the trade called GUTs. As the plurality of the name suggests, there are several versions of these theories. (The prototype theory was published by H. Georgi and S.L. Glashow in 1974.) Because the theories treat on the same footing the fields of leptons and quarks, in principle they all allow, in principle, for particle interactions that are not observed, some of

years for this form of decay).

That creates a dilemma for the exponents of the GUTs. May it be, they ask, that the reason for these and other difficulties is not that unification is an impossible dream, but that the attempts so far made yield awkward predictions precisely because unification is still incomplete? Maybe it would be better if gravity were included?

That has inspired the enthusiasm, since the early 1980s, for radically different

which then have to be excluded by restrictions of the kinds of field allowed (as with the device for excluding right-handed neutrinos from the electro-weak theory).

Even so, some unfamiliar phenomena seem unavoidable, the chief of which is the prediction by GUTs that the proton should be capable of decay not into a neutron, which would be energetically impossible, but into leptons. The first predicted lifetime of the proton against these decays, 10^{31} years, has been ruled out by measurements by the Italian-led group operating the Gran Sasso particle detector beneath Mont Blanc (which has already fixed an upper bound of 10^{33}

ways of dealing with the problems of interacting fields. One of these is what is called 'string theory', in which the intrinsic energy (the mass) of the fundamental particles is taken to be the equivalent of the energetic excitation of a loop of matter, a closed string. Excitations of such a structure may include, for example, stationary vibrations.

The attractiveness of this way of looking at the particle problem is that it seems to offer a way of dealing with the gravitational interaction between matter particles in the same way as the three other interactions. Part of the price to pay for this advantage is that the strings must be supposed to exist in at least ten dimensions. Calculations are carried out in the full range of dimensions, but then six of the dimensions are dispensed with by supposing that they are 'compactified', which means that they are wrapped around on themselves in loops whose dimensions are related to the quantum scale of the phenomena that the structures describe.

String theory has been grandly elaborated in the past decade, but has not yet established itself as the favourite. But it has raised a question of some philosophical interest, more often discussed over laboratory lunches than in formal publications: is it possible that the world we live in is, in some sense of the word 'real', really a ten-dimensional world, and that our view of it is conditioned by the limitations of our perceptions of it to merely three space dimensions?

Meanwhile, the search for unification will continue. The prize is too great for people to leave alone. How else will it be discovered why (rather than how) there appear to be just three generations of quarks and leptons? There is a sense in which the question "Why is matter like it is?" remains entirely open. □

Bad press, bad luck

PARTICLE physics has had a bad press (and bad luck) in the past year or so. The explanation is the high cost of the next generation of particle accelerators.

The cancellation a year ago by the US Congress of the \$8-billion-plus Superconducting Super Collider (SSC), on which construction had begun in Texas, has been a body-blow to the experimenters. Continuing uncertainty over the budget for CERN's planned Large Hadron Collider (LHC) further casts them down.

Have the world's politicians implicitly decided that the origin of matter is an unimportant question? Or one whose answer would not have the interest to justify the cost?

That would be a great misfortune.

Nature's consistent view is that the search for an understanding of why matter is as it is compares in interest and cultural importance with all the other grand questions still unanswered, but that the nature of the question is such that only a genuinely international organization can successfully ask for the funds required.

The immediate need is to understand the role of the Higgs boson (see page 23) in the standard model. Beyond that lies the question of whether there is particle physics beyond that apparently complete, but still imperfect, view of the world.

When the whole history of particle physics has been such a succession of surprises, each of which appears to have uncovered an unsuspected layer in a

multi-layered intellectual structure, and when the most recent elaboration of the theory (QCD) is merely 20 years old, would it not be great good luck if present understanding were complete?

Without an accelerator beyond the LHC, particle physics will in part be driven back to where it began with cosmic rays, to the study of celestial events (see page 27). When more is known of quasars, the accessible natural laboratories may be multiplied in number. There are also the recently discovered γ -ray bursts that point to energetic events somewhere in the Universe.

Terrestrial laboratories also still have much to offer. The first confirmation of parity violation in the weak interaction came from a laboratory experiment, after all. 'Small Big Science' is not yet over. □

Stars, planets and the Earth

It is puzzling now to be reminded that the Lowell Observatory at Flagstaff in Arizona was founded exactly a century ago because of the enthusiasm of Percival Lowell for the study of Mars and of its presumed inhabitants. But that is simply a measure of the degree to which observation of the Solar System

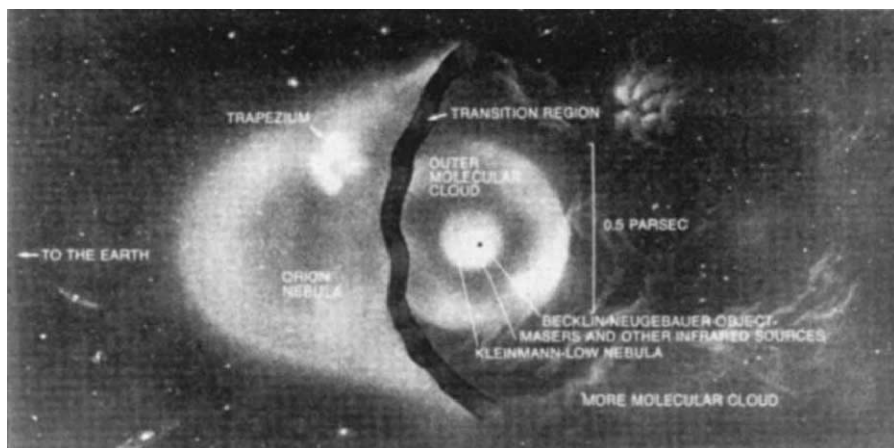
between stellar size and luminosity, or between luminosity and colour (now known as the Hertzsprung–Russell diagram), was recognized only during the First World War. In the light of Lane's hypothesis about the generation of energy by gravitational contraction, it was inevitable that the direction of evolution

by-product of stellar modelling was the argument in 1957 by Geoffrey and Margaret Burbidge, W.H. Fowler and F. Hoyle that the nuclei of all but the lightest elements (^2H and ^3He) are synthesized in stars, with the unavoidable conclusion that many stars are made from the debris of others belonging to earlier generations. (The observational proof of that is that the stars of the Galaxy's globular clusters, the oldest that there are, are uniformly deficient in heavy elements relative to the Sun.)

That sequence of developments has transformed the study of the other objects in the Galaxy (and, by extension, of the stars of other galaxies). In principle, there is no lacuna to impede the understanding of any Galactic object that may be found in a telescope image. In practice, of course, not everything is that simple (see below).

Nor do we yet fully understand the Galactic objects closer to home — the planets of the Solar System and their attendant satellites. That the Solar System was formed by the collapse of a cloud of dust and gas, much as is now happening in the Orion nebula (see figures) is easily accepted. That there should be gradation of the properties of the planets from Mercury (nearest the Sun) to Neptune (the most distant unexceptionable planet) is also understandable.

But it is not as yet crystal-clear whether the gradation is better explained by the aggregation of nebular material in planetary orbits followed by the loss of light molecular constituents under the influence of the solar wind, or by the prior condensation of refractory materials in the orbits of the inner planets, most probably into objects a metre or so in diameter in the first instance. (The second is the more widely accepted view.)



Schematic diagram of the structure of the Orion molecular cloud. See below for Orion's appearance in the sky. (Kindly supplied by T. L. Wilson of the Max Planck Institute for Radioastronomy.)

has been made meaningful in the intervening century.

Drawings of the surface of Mars had revealed patterns that changed from time to time. (The first editor of *Nature*, Sir Norman Lockyer, vigorously contributed to the quest for data.) At the end of a century distinguished by major civil engineering exploits on the surface of the Earth, would not the copernican principle naturally prompt the hypothesis that similar undertakings were under way on Mars? And because the observed scale of the operations was greater than any known on the Earth, did it not follow that the Martians were "more intelligent"?

The explanation of the confusion at the end of the nineteenth century about Mars and its inhabitants almost certainly stems from the seasonal variation of the covering of solid carbon dioxide ('dry ice') over the surface, the differential transparency of the atmosphere with colour and the pronounced linear structures now seen in spacecraft photographs of the Martian surface.

At the beginning of this century, there was also great confusion about the evolution of stars, first made plausible by the US mathematician J.H. Lane in the 1870s. Lane seems to have been the first to argue that the gravitational collapse of a mass of gas would lead to an increase of the temperature — the only mechanism of stellar energy generation until the recognition in the 1930s that the fusion of light nuclei was probably the source of energy.

The regularity of the relationship

should be mistaken. Red giants were supposed to represent the early stages in the process of collapse and not (as now) the expanded condition of stars that have exhausted their hydrogen fuel and are burning helium.

In the circumstances, it is remarkable that so much was so quickly accomplished, in the decade or so after Bethe's first account of the thermonuclear energy source, to put stellar modelling on an objective and quantitative basis. A crucial

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

The Orion Nebula M42. Taken from *The Cambridge Atlas of Astronomy* edited by J. Audouze & G. Israël. Cambridge University Press, 1988.

Whatever the truth, the marked difference between the external skins of Venus and the Earth is a challenge to credulity, given that the two objects are similar in

now appears to be that the satellite is a significant conductor of electricity. This effect is enhanced by the presence of ionized material in a doughnut-shaped

NASA

ring broadly following Io's orbit. Sulphur ions in this 'plasma torus' seem to originate from Io's prolific volcanoes, which erupt enough sulphur and silicate lava to renew the satellite's surface every century or so.

Present knowledge of the Solar System is almost entirely the product of observations from instrumented spacecraft in the past three decades. One has the impression that data are being gathered more quickly than they can be assimilated.

In many ways, the same is true of present understanding of the structure of the Earth, which is not to diminish the importance

of the great achievement of the past quarter of a century in providing a coherent and persuasive account of the tectonic movements of the Earth's crust.

Wegener's turn-of-the-century conjecture about "continental drift" has been amply and, in the circumstances, quickly confirmed by the inference of sea-floor spreading from the magnetic lineations on the basaltic ocean floor, by the demonstration of past (and present) movement of the land-masses by palaeomagnetism (and more recently by very-long-baseline interferometry), by the seismic recognition of inhomogeneities in the Earth's mantle (as at subduction zones, the places where descending

oceanic plates re-enter the mantle) and by conventional geology.

The difficulties with the Earth are, now, more subtle. The driving force for plate tectonics is some form of convection in a solid upper mantle which nevertheless behaves as a fluid over long timescales, perhaps because it is inhomogeneous on some length-scale. And the world awaits a convincing illustration of why upwelling material takes the form observed at mid-ocean ridges and rift valleys, as well as of the way in which lateral transport of material contributes to the force on the underside of spreading sea floor. (The negative buoyancy of descending plates is another impetus.)

The deeper interior of the Earth is another source of controversy. Is the discontinuity at ~660 km depth a consequence simply of a pressure-induced phase change, or is there also a chemical discontinuity? (Experiments with rock material at high pressure, coupled with laboratory and numerical simulations of mantle convection, should eventually resolve this question.)

Even the molten outer part of the Earth's (predominantly) iron core is now accessible to simulation in the laboratory, although the solid inner core is still out of reach. But the identity of the light, alloying element(s) in the core is still unknown, as is whether the molten core is reacting with the silicate mantle above, as well as crystallizing at the boundary with the inner core below. These questions, still unanswered, are important not simply because we inhabit the Earth but also because an understanding of the other inner planets will be more easily won when the Earth is fully understood. It is, after all, likely to be more accessible for some time to come. □

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Voyager spacecraft image of Jupiter's moon Io. The large dark spot in the centre foreground is Pele, one of Io's active volcanoes.

mass and similarly placed relative to the Sun. That the atmosphere of Venus should be rich in SO₂, and thus optically opaque, is another puzzle, as is the topography of the surface as determined by radar measurements. Are the differences properly accounted for by nothing more than the circumstance that the surface temperature of Venus exceeds 200 degrees Celsius?

The satellites of the major planets are a puzzle in their diversity. That there is something peculiar about the Jovian satellite Io was first apparent in the 1960s in the coincidence of the bursts of decametric radio emission from Jupiter with the rotation of Io in its orbit; the explanation

Laboratories for the future?

NEUTRON stars have a curious history, not least because there was a sense that they should exist three decades before they were found in the real world.

The now-famous tale is that S. Chandrasekhar, a graduate student at Cambridge in the early 1930s, set out to calculate the stability of a star like the Sun as the hydrogen in its core was depleted (by conversion into helium) by thermonuclear reactions. The production of energy from hydrogen becomes less efficient as the fuel is diluted with its waste-product.

Several things can then happen, one of which is that a very-low-mass star will sedately shrink, becoming a white dwarf whose envelope is supported by the pressure of the degenerate electron gas formed as electrons are squeezed out of

the atoms to which they belong. In higher-mass stars the core collapses, but its temperature increases so that it can burn helium instead of hydrogen, the envelope expands, and the star will be a red giant — "red" because there is vastly more surface to radiate away an admittedly increased energy production.

What happens next, as the star reaches the end of its luminous life, is more dramatic still. The envelope of a star like the Sun is supported by the flux of radiation from the hydrogen-burning core, but the timescale for the transport of energy through the envelope may exceed a million years. If the collapse of the core is faster than that, the envelope may be thrown off. (The Galaxy is replete with such 'planetary nebulae', which have nothing to do with planets.) The highest-mass

stars will explode as supernovae, which may result in a white dwarf or a neutron star. Interestingly, stars that follow the route to red giants are expected also to end up explosively, as supernovae; at some stage the nucleosynthesis reactions in the core will produce energy more quickly than it can be carried away by the envelope, which will then also be blown away.

Despite the expectation that they should exist, the discovery (in 1967, appropriately at Cambridge) of pulsating radio sources was a complete surprise — even to the authors of the discovery (the late Sir Martin Ryle's group). Since then, neutron stars (recognizable by their radio pulsations only) have become invaluable pointers to the characteristics of astrophysical phenomena.

Binary pulsars (in partnership with other neutron stars or white dwarfs, or even with black holes) promise to provide

The technical dimension

BEFORE the Second World War, there were just two windows on the world beyond the atmosphere of the Earth: the radiation from external stars and galaxies that happens to fall within the part of the electromagnetic spectrum visible to the human eye, and cosmic rays. In practice, the study of cosmic rays was chiefly of interest as a route to particle physics without particle accelerators. The detailed study of the ionosphere was a way of gathering information by proxy about the behaviour of the Sun, and was not especially successful.

The transformation of astronomy since 1945 consists of the extension of astronomical observations to what is essentially the whole of the electromagnetic spectrum. The development of radio-astronomy in the 1940s and 1950s appears to have been inspired directly by the successful wartime use of radar based on microwaves, with wavelength of the order of 1 cm or so.

Similarly, the development of infrared astronomy, beginning in the 1960s, owes much to the development of sensors for military purposes, while Herbert Friedman, one of the pioneers of X-ray astronomy in the United States, relied on captured V2 rockets to launch his first X-ray experiments.

But the refinement of these techniques in the hands of astronomers has yielded accuracy and sensitivity of a kind to which the military would not have aspired.

Much the same characterizes the exploration of the objects of the Solar System in the past few decades. Here on the



FIG. 1 Simulated view of the surface of Venus, constructed from radar elevation data obtained by the Magellan spacecraft. The main feature is Gula Mons, a 3-km-high volcano.

Earth, radar was used in the 1960s to provide a picture of the solid ground beneath the Antarctic ice cap, just as it has been used by the US Magellan spacecraft to provide a detailed (and astonishing) picture of the topography of the surface of Venus (see Fig. 1).

Military and diplomatic interests also provided, in the 1960s, a powerful impetus for the development of sensitive seismographs for the detection of nuclear explosions, which has powerfully reinforced the traditional use in geophysics of seismic signals from distant earthquakes as a means of inferring something about the internal structure of the Earth. The more recent spectacular results of the use of sonar for the exploration of the deep-ocean floor is a similar case (see Fig. 2).

To refer to these connections is not to diminish astronomy or geophysics, but rather to raise a question about the future. If, with the ending of the Cold War, military development budgets are everywhere reduced, will basic science have to rely on its own resources, themselves ultimately provided by governments, for the continued development of techniques? Or is it realistic to hope that the partnership with industry that governments everywhere are urging on the research community will sustain the development of new techniques?

Perhaps the most remarkable development has been the widespread use of the light detectors known as charge-coupled devices (CCDs) — solid-state photon detectors whose sensitivity, at least in the visible part of the spectrum, approximates to the ideal, and the response of which (unlike that of photographic emulsions) is linear with intensity. CCDs have become in 15 years a standard source of images of the sky, and one pre-digitized for easy computation. □



FIG. 2 Shaded-relief image of the East Pacific Rise and its intersection with the Clipperton transform fault. High-resolution 'swath'-mapping using sonar systems such as Sea Beam has provided a detailed view of mid-ocean-ridge structure and magmatism.

Plotting software courtesy of R. Tyce, NECOR/Sea Beam facility

some of the most stringent tests of general relativity, as well as ways of verifying the mechanisms of supernova explosions.

But there is also a reasonable chance that they may provide a means of testing the predictions of the standard model of particle physics, admittedly at the cost of turning experimental physics into observational science.

The density at the core of a neutron star with a mass equal to that of the Sun may exceed $3 \times 10^{18} \text{ kg m}^{-3}$, or 3 billion tonnes per cubic millimetre. The forces responsible for the confinement of this material are the familiar forces of gravitational attraction. But the density is so great that the separation of electrons from their atoms believed to be responsible for the density of white dwarfs is taken a step further: electrons combine with protons in a process which is the inverse of the usual β -decay of the free neutron.

The result is that most of the material in a neutron star with a mass comparable with that of the Sun will consist of free neutrons, mixed with free protons and electrons. A stable neutron star is then

prevented from further collapse under self-gravitation by the way in which the nucleons (which are fermions) form a degenerate Fermi gas, exerting an outward pressure. In these extreme conditions, the nucleons are expected to be superfluid — and, in the case of the protons, superconducting.



M1, the Crab nebula, is the expanding remnant of a star that was seen to explode almost 1,000 years ago. It now hosts the product of that explosion — a pulsar.

Taken from *A View of the Universe* by D. Malin (Sky Publishing/Cambridge University Press, 1993)

Like other astrophysical objects, stars and planets for example, neutron stars have a layered structure. Outside the degenerate central region is one in which still intact nuclei are mixed with superfluid neutrons. The outer crust of a neutron star, with a density upwards of $10^{10} \text{ kg cm}^{-3}$, is believed to be essentially a conducting solid. To complicate the physical situation, pulsars appear to have magnetic fields exceeding 10^9 tesla — which is why they are detectable by radio pulses synchronous with their rotation.

Part of the interest of the observable neutron stars is that they are potentially laboratories for the study of a variety of phenomena, not all of them astrophysical. E. Witten, for example, has argued that the inner core of such a star may consist in part of 'quark matter' — a state in which quarks are no longer confined to individual hadron particles.

There have already been speculations in the literature about the geometrical shapes that would be occupied by quark matter in the core of a neutron star. Direct observation of such arrangements will clearly be impossible, but there is a sporting chance that the study of neutron stars over extended periods of time will make it possible to gather an understanding of the properties of neutron matter at high pressures that will be of value in itself, and which will also complement and extend the information that can be gathered from the particle accelerators now likely to be built. Pulsating stars have come to stay. □

How did the Earth form?

THE Galaxy abounds with objects whose nature and origin is poorly understood, from the X-ray binary stars in which mass is being transferred from one to another to the organic molecules accumulating in molecular clouds. Here is a set of problems much closer to home: how did the Earth form, 4,500 million years ago (± 50 million years)?

The question is important not only for its own sake, but because the Earth and the other inner planets are outwardly solid objects, unlike Jupiter and the other much larger objects that lie beyond them; they must therefore have something significant to say about the evolution of the Solar System as a whole, as well as providing a guide in the serious search there will eventually be for habitable planets elsewhere in the Galaxy.

Perhaps significantly, the oldest rocks are dated (by radioactive clocks) to 3.8 billion years. That date fortunately corresponds with that of the second copious wave of meteoritic impacts on the surface of the Moon, one of the few substantial

data won by the Apollo programme, suggesting that the Earth accreted a significant but unknown fraction of its mass in the same process.

What was there (or here) before that? There are two lines of evidence — calculations of collapse in the supposed solar nebula and observations of other early star systems. The rarity of the latter accords with what is known of the relative ages of the Sun and the Earth, which suggests that the inner planets were formed within a few hundred million years of the Sun.

The central problem in the formation of any star is the mechanical problem of dispensing with angular momentum during the collapse of a cloud of gas and dust. Calculation and observation agree that the task is usually accomplished either by the formation of a double star or of a planetary disk perpendicular to the axis of the star.

The favoured view of the formation of the inner planets hangs on the condensation of oxides of elements such as iron and silicon in the inner regions of the solar

disk. The time course of the evolution of temperature in a collapsing proto-star is not well enough known to reconstruct with confidence the likely course of events during the collapse. Would the temperature in the solar disk have been high enough to vaporize pre-existing dust grains? And what would have been the cement that held together the condensing grains so as to allow them to grow to the solid objects of the order of 1 metre?

The aggregation of planetesimals into more substantial objects the size of planets is another imponderable. It is possible that this process began only when the formation of the giant planets, and Jupiter in particular, provided perturbing forces sufficient to break up into clumps a system of planetary rings much like that now seen around Saturn, with the difference that the constituents were minerals (not ice) and the central object was the Sun.

So is it possible that the second wave of meteoritic impacts on the surface of the Moon 3.8 billion years ago represents the last recorded stage in this enforced aggregation? □

Which is the chicken, and which the egg?

THE reductionists' agenda cannot be complete until we know, in general terms, how life began. Nor will it be possible to conduct a meaningful search for where else in the Galaxy life might be found without knowing more about the circumstances under which terrestrial life began.

Unfortunately, there is as yet not much that is tangible to report. More accurately, there have been many pointed investigations since the Second World War, many of them with great intrinsic interest, but there is not yet an unambiguous pointer to the mechanisms that may have led to the emergence of living organisms in the fossil record of 3.5 billion years ago.

Some features of the problem to be unravelled are nevertheless clear. First, the atmosphere of the early Earth was probably a reducing atmosphere in which molecules such as methane (CH_4) and ammonia (NH_3) would have been prominent. That belief rests on the predominance of hydrogen among the constituents of the solar nebula and on the observation that present-day eukaryotic cells require detoxifying systems for their survival.

But on the common assumption that the first self-replicating organisms would be unicellular, and that they would also have been photosynthetic, oxygen would have had to be present in the early atmosphere as CO_2 . Experiments by Harold Urey and Stanley Miller at Chicago in the early 1950s, in which electric discharges were passed through mixtures of CH_4 , H_2 ,

NH_3 and H_2O , yielded small quantities of the simpler amino acids, suggesting that they may have been produced from the early atmosphere.

A second general principle in the search for the origin of life is that subsequent events must have followed an evolu-



Light micrograph (magnification $\times 94$) of a thin section of Gunflint chert from Ontario, Canada, showing fossil remains of the first Precambrian life-forms to be discovered, dating from two billion years ago. The sample includes a mixture of organic detritus, filaments resembling the modern day blue-green algae, and colonies of algae (spikelets). These primitive plants gained energy through photosynthesis, releasing free oxygen into the environment as a by-product. This had far-reaching consequences for the subsequent history of the environment.

tionary path selected from among many alternatives by environmental circumstances. For that reason (among many others), there has recently been great interest in what is called 'molecular evolution' — the use of artificial selection pressures to select products with particular properties from systems of chemical reactions using random mixtures of molecules as substrates. Their relevance to the origin of real life is to demonstrate that selection does, unsurprisingly, function at the molecular level.

Much also can be (and has been) learned from the presence in extant organisms of macromolecules with partic-

ular functions which are also highly conserved in their structure across species. Thus, on the basis of comparisons of ribosomal RNA (rRNA) species, C. Woese has argued that extant unicellular organisms can be divided into no more than three groups. But the inference that the elements of rRNA structure that are common to the three groups are truly primaevial is not unavoidable; the common elements could have evolved in parallel, in changing but common circumstances.

That is not to say that the biochemistry of extant life-forms has nothing unambiguous to say about the origin of life. On the contrary, there are many features of the present diversity of life that could (and should) be more systematically exploited to create a chronologically based taxonomy of extant organisms. But it is unlikely that molecular taxonomy in itself, invaluable though it has proved to be in demonstrating the evolutionary relationships between closely related organisms, will be sufficient to identify a set of biochemical reactions that could be regarded as the minimum requirement of primitive organisms.

One speculation, for example, is that the long gap of about two billion years between the first appearance of micro-organisms in the fossil record and the rapid radiation of multicellular organisms in the early Cambrian period was spent in the evolution of the intercellular signalling systems and the genetic mechanisms of cell differentiation necessary for the functioning of multicellular organisms.

The obvious difficulty is that the genes responsible for the intercellular-signalling molecules (and their receptors) in primitive organisms are probably adaptations of pre-existing genes with a continuing role in the metabolism of cells. Even so, a more systematic and extensive comparative study of extant organisms which appear on general grounds to have links with the primaevial should eventually yield some kind of picture of what the most primitive unicellular organisms were like.

Sadly, that is not the origin of life. □

RNA in primaevial genetics

SINCE the discovery in the early 1980s that RNA molecules may have catalytic properties (as in the splicing of messenger-RNA molecules in cell nuclei), there has been a revival of interest in the possibility that RNA was the Earth's first self-replicating molecule, although the 'RNA world' was first inspired by the high degree of conservation of the sequence structure of ribosomal RNA (see above).

There are three arguments in favour of RNA as the first biomolecule¹. First,

being a polymer of alternative nucleotides, it can carry genetic information (as if it were DNA). Second, there is laboratory evidence that the polymerization of HCN (probably a constituent of the primitive atmosphere) can lead to the formation of purines, including two of the components of modern RNA molecules, adenine and guanine. (The hydrolysis of polycyanogen can have similar results.) Third, simple chemicals related to nucleotides are ubiquitous in extant life-

forms, as coenzymes and also as ATP.

Unfortunately, there are also difficulties, not least the source of the pyrimidine nucleotides that are components of RNA, cytosine and uracil, as well as the source of the ribose with which they are specifically (and stereochemically) linked in modern RNA. The problem with the sugars is not the likelihood of their occurrence on the primitive Earth (formaldehyde, CH_2O , is a potential source) but that of specificity; why the link with ribose, when other sugars may have been equally abundant?

Several ways of circumventing those

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

Scanning electron micrograph of a spore tower of the slime mould *Dictyostelium discoideum*. Slime moulds are a curious division of plants with many of the characteristics of animals. Magnification $\times 140$.

difficulties have been suggested, notably that the original backbone of a polynucleotide may have been related to glycerol rather than ribose, or that the polymerization of early RNA molecules was guided by a structured solid surface, perhaps a clay. But in the absence of a successful demonstration that systems of that kind would function as it is inferred they did, the question of how the RNA world began

remains unanswered.

On the assumption that the starting materials for RNA synthesis (nucleosides, or nucleotides linked to a phosphorylated sugar) come into being, their autocatalytic replication seems plausible enough. G. von Kiedrowski has demonstrated² the autocatalytic self-replication of an analogue of a hexanucleotide, using trinucleotides as starting materials. J. Rebek has described a somewhat different scheme³.

L. Orgel at the Salk Institute has demonstrated the successful copying of polynucleotide analogues up to 14 units long⁴, but in a system that is unlikely to have been the one used on the early Earth because of the readiness with which the molecules cyclize and otherwise wrap up on themselves.

More than two decades of the pursuit of self-replicating RNA systems, which has brought a rich harvest of information about the conditions under which replication may occur, has nevertheless also left a sense of disappointment; no artificially devised system has proved to be simple enough to be plausible, and robust enough to suggest that, on its own, it could have triggered the emergence of life.

Some of these difficulties may be sur-

mounted by more recently suggested (and demonstrated) replication devices. In May this year, for example, it was shown⁵ that 24-nucleotide DNA molecules could be replicated, without the intervention of enzymes, in the major groove of a DNA duplex molecule serving as a template. It remains to be demonstrated that the scheme will work for molecules of general sequence, while the use of DNA duplex molecules supposes a big step beyond the RNA world.

One obvious difficulty with studies of this kind is their inevitable assumption that the replication of informational molecules on the primeval Earth was self-contained, and unassisted by other chemical components of the primeval soup. That is a stringent assumption, and probably incorrect. Even so, studies of this kind are likely to continue, not only for their own sake but because of the vivid interest of the chemical industry in ordered copolymers, which may have many advantages over existing polymer materials. □

1. Joyce, G.F. *Nature* **338**, 217–224 (1989).
2. von Kiedrowski, G., Wlotzka, B. & Helbing, J. *Angew. Chem. int. Ed. Engl.* **28**, 1235–1237 (1989).
3. Hong, J.-L., Feng, Q., Rotello, V. & Rebek, J. *Science* **255**, 848–850 (1992).
4. Orgel, L. E. *Nature* **358**, 203–209 (1992).
5. Li, T. & Nicolaou, K.C. *Nature* **369**, 218–221 (1994).

Guesswork and randomness (for now at least)

WHAT came before the RNA world? That the first living things would not have been recognizable as such is readily accepted. They may simply have been aggregations of organic molecules which, under the illumination of sunlight, allowed the evolution of particular species, RNA molecules perhaps.

Decades ago, J. D. Bernal suggested that such processes might be sustained on the surfaces of clay particles. A newer version of that picture, due to Günter Wächtershäuser, is that the surfaces would have been minerals such as pyrite, capable of binding and polymerizing simple organic molecules fashioned from five-carbon isoprene units carrying phosphate and hydroxyl substituents¹.

One virtue of the model is that, with the conversion of polymerized isoprene units into glycols, simplified cell membranes could be formed along with accumulating chemicals. Wächtershäuser describes an intermediate stage of evolution in which membranes would enclose solutions of evolving chemicals between themselves and the solid substrate beneath.

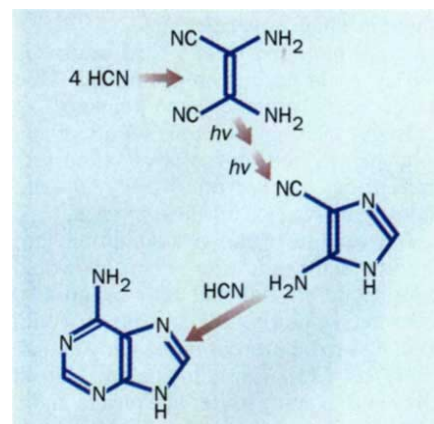
More recently, the argument has been carried further by the identification of terpenoid molecules from petroleum and

other geological sedimentary deposits with chemicals known to be constituents of extant cell membranes².

So what was the actual course of events more than 3.5 billion years ago? Detailed reconstructions of the emergence of RNA, or alternatively of isoprene-based chemicals, becomes guesswork of a kind.

So why not guess from the outset? That is the basis of the now common interest in 'molecular evolution', partly stimulated by an experiment of Spiegelman in the 1960s. The RNA gene for the replicase enzyme (a protein) of the *Q β* bacteriophage was used repeatedly to replicate further copies of the viral RNA in several cycles, generating departures from the original sequence in the process. Selection for the replicase giving the most rapid replication of the whole virus eventually yielded a dramatically shorter enzyme.

Now that the production of random sequences of RNA is relatively simple (by producing random sequences of DNA in a synthesizer, and then converting them to RNA, for example) there have been demonstrations of how to evolve, with the help of a selection process, RNA molecules with particular properties — their binding to small molecules³ or their



Synthesis of the nucleotide adenine (bottom left) from HCN.

enzymatic function as ribozymes⁴.

The prophet of this approach is Stuart Kauffman, from the Sante Fe Institute, New Mexico. A patent originally in his name is now owned by the Darwin Institute in Seattle, Washington. Expectations among industrial molecular biologists are high. Whether the same approach will be applicable to real life is another matter. □

1. Wächtershäuser, G. *Microbiol. Rev.* **52**, 452–458 (1988).
2. Ourisson, G. & Nakata, Y. *Chem. & Biol.* **1**, 11–23 (1994).
3. Ellington, A.D. & Szostak, J.W. *Nature* **346**, 818–822 (1990).
4. Bartel, D.P. & Szostak, J.W. *Science* **261**, 1411–1418 (1993).

The age of australopithecines

How can we hope to understand our place in nature until we can understand how we have come to occupy our present niche on the surface of the Earth? That question is a sufficient case for the more energetic pursuit of an understanding of human evolution. Yet, as things are, there is only a skeletal account of how it happened.

Luckily, the flowering of molecular genetics has now provided palaeoanthropologists with a temporal framework within which to fit the scraps of fossil evidence they recover. The hope must be that further understanding of the human and related genomes will make it possible to ask sharper questions about human evolution than are now in people's minds.

Darwin's view that people share a common ancestor with the extant great apes was based on behavioural as well as anatomical evidence, but there was then insufficient evidence with which to construct a timescale for the process. Now, the molecular and fossil evidence are largely in agreement: on the basis of the accumulation of mutations in comparable genes, the separation is put at between 6 and 7 million years ago. The emergence of modern *Homo sapiens* is dated at about 200,000 years ago on the basis of measured variations of the sequence of human mitochondrial DNA¹.

The second date is the more striking. It means that the history of *H. sapiens* spans only about 10,000 generations, making people one of the youngest of substantial species on the surface of the Earth.

The first date is also consistent with the fossil record. The intermediate fossils

assigned on morphological grounds to species of the genus *Australopithecus* are, as they should be, all much younger than 6 million years. The significance of this year's find² in Ethiopia, called *A. ramidus*, is that it appears to represent a genuine australopithecine with an age of 4.4 million years, helping to fill the gap between the supposed divergence of the human and ape lineages and the oldest fossil australopithecine previously known, *A. afarensis* at 3.9 million years¹.

Although *A. ramidus* is judged, both by the authors of the discovery and by others, to be more closely related to the chimpanzee than to the gorilla, it seems unlikely that palaeoanthropologists will consider that question to be decided by the handful of teeth and the fragments of cranium now recovered from Ethiopia. The best hope is that this winter's expedition will return with more definitive material.

But the discovery of an authentic australopithecine at 4.4 million years does at least provide palaeoanthropologists with an opportunity and an incentive to tidy up the more recent evolution of these species towards *H. sapiens*. On the assumption that the successive species of australopithecines now recognized must include a lineage of continuing selective adaptation,

it should at least be possible to exclude some of them from consideration as ancestors of *H. sapiens*.

One casualty of this process may be that the diverse group of fossils from 1 million years or so ago, known as *H. habilis*, may be more properly recognized as australopithecines. But there is no whit of evi-



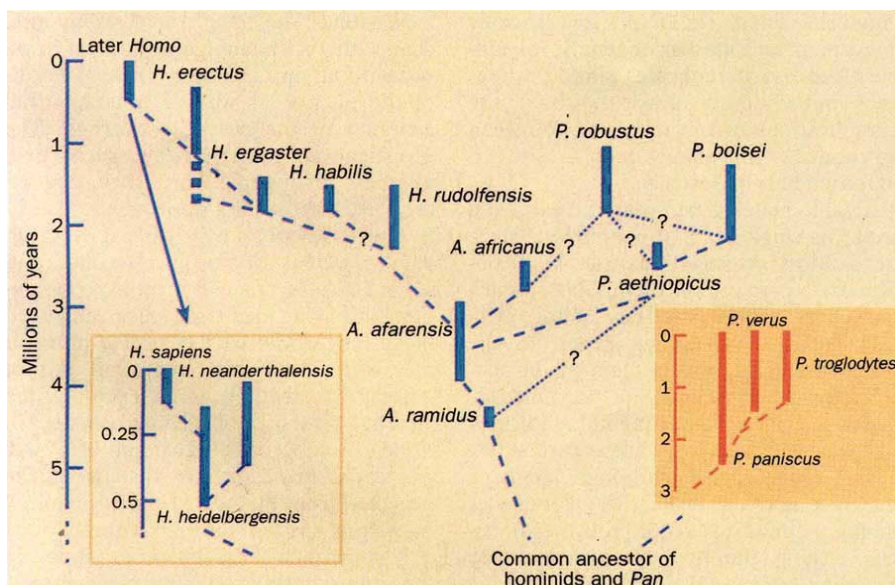
Fossil of a child's jaw from Aramis in Ethiopia, one of a series representing the earliest hominids in the world.

dence discordant with the conclusion that the adaptation leading to *H. sapiens* took place in Africa. The outward radiation of the species to the rest of the world is another and more speculative matter.

One of the frustrations of palaeoanthropology is that the australopithecines are better represented by fossils than are the great apes with which they were, presumably, contemporaries. This is an important question, not merely because it is at present unclear with which of the two extant species of *Pan* the modern *H. sapiens* should be compared, but also because a comparison of the parallel evolution of the African great apes would throw light on the environmental influences that may have helped to shape the adaptation of both lineages.

On the face of things, the importance of environmental influences seems not to have been great, but that may simply mean that too little is known of the past climate and ecology of Africa. It is relevant, however, that the separation of the human and great ape lineages long predates the onset of the Pleistocene period, during which ice covered many regions at high latitude, and when it appears that rainfall in Africa was greater than now and in the preceding Pliocene period.

It may, of course, have been that the evolution of the australopithecines was driven not by external influences but by the selective advantages of mutations that, arising within a genome, confer on those who live by it either decisive advantages over competitor species or



Provisional hominid evolutionary tree. The vertical bars show the known dates of first and last appearances of species. The most likely relationships between hominid taxa are reflected by the dashed lines; dotted lines reflect less likely alternative relationships. (*Nature* 371, 280–281; 1994).

unusual adaptability.

The high interest of this question is that *H. sapiens* differs from the great apes (including the orang-utan of Southeast Asia, no longer counted as a close relative) in having 46 rather than 48 chromosomes. The genes of the missing ape chromosome have now been identified as a single block located at one end of human chromosome 2. It follows that at some stage in the evolutionary process, there was a major chromosomal translocation, as a consequence of which an entire ape chromosome was incorporated into the human chromosome 2.

It is not unreasonably soon to ask when, in human evolution, this translocation occurred, and even to guess what its genetic consequences would have been.

On the face of things, it would explain the lack of correlation, in human history, of adaptations with external events, but this may be to say no more than that the evidence so far of the environment in which the australopithecines evolved is entirely insufficient to make a judgement. That is yet another reason for hoping that the study of the comparative evolution of apes and people will soon be possible.

Meanwhile, it is possible that something could be learned from the study of the genomes concerned. What are the genes translocated onto human chromosome 2? Is there any reason to believe that their presence there is advantageous, perhaps because it is plausibly linked with the known progressive increase of the size of the brain-case of the australop-

ithecines, and thus presumably with a progressive increase of the sophistication of the nervous system?

That is one way in which the human genome projects may assist the cause of palaeoanthropology. Another will be possible when the nucleotide sequences of the non-coding regions of the human and ape genomes are known. But a great-ape genome project (for which there appear to be no present plans) would be of great interest in its own right, if only as a means of cataloguing the inter-specific differences between two closely related lineages, at least one of which has evolved at remarkable speed. □

1. Cann, R. L., Stoneking, M., Wilson, A. C. *Nature* **325**, 31–36 (1987).
2. WoldeGabriel, G. et al. *Nature* **371**, 330–333 (1994).

Our ancestors' ancestors

WHATEVER the uncertainties remaining about the course of evolution of the australopithecines, the question of where their ancestors came from and what they were like is even more obscure. Technically, they are known as hominoids. As a group, they include the great apes and human beings (the Hominae), the Ponginae (the orang-utan and its presumed ancestors) and the predecessors of extant creatures such as gibbons¹.

Among the common attributes of hominoids are the flexibility of the forearms and their components — the elbow, wrist and so on. There is unambiguous evidence from Africa that the hominoids were established there 20 million years ago, alongside other primates, where they are represented by species of the genus *Proconsul*.

But what then is to be made of the unambiguous hominoids represented by fossils from Pakistan and elsewhere, with ages of the order of 10–20 million years? The issue is controversial. The simplest view is that they should be grouped with the genera such as *Sivapithecus*, amply represented by fossil specimens, and regarded as ancestors of the orang-utan, now exclusively Asian. But that is widely disputed.

Whatever the correct interpretation, these hominoids appear to share with the early australopithecines the flexibility of the forearms and the woodland or forest habitat that appears to have been the common cradle of human emergence. The trick of enabling the thumb and forefinger to act in concert is common among them. The anatomical evidence suggests that they could walk with the help of their forearms, but that bipedalism awaited the emergence of *Homo erectus* in the human lineage. □

1. Andrews, P. *Nature* **360**, 641–646 (1992).

Migration out of Africa

IN 10,000 generations or thereabouts, human beings have migrated from Africa to populate almost the entire surface of the Earth. It is known that there were people in Australia 50,000 years ago, and that the Americas were occupied less than 25,000 years ago by land migration from Asia. But there are limits to what can be learned by archaeology. Yet the cultural history of *H. sapiens* is presumably faithfully recorded in the genes of people still alive. Unfortunately, the attention given to deciphering these records does not match the intrinsic interest of a successful outcome.

Most attention has so far been given to the genes of mitochondrial DNA, largely as a result of the urging (and the example) of the late A.C. Wilson (Berkeley). Mitochondrial DNA (mtDNA) is especially convenient as a marker of genetic relatedness because it replicates simply, whenever mitochondria divide, without the complications of recombination between autosomal chromosomes, and because it is inherited only maternally.

Usable material can often be extracted from the single hairs of people long since dead. Most investigations use the 1,000-plus base-pair region of mtDNA (which has nearly 17,000 base-pairs altogether) that includes no coding genes, in the expectation that most of them will be subject to mutation without affecting function.

Language is another possible indicator of ancestral relatedness, but is intrinsically a less reliable one. Although language may be one of the distinctive attributes of human beings, it is acquired by newborns from their parents. Nevertheless, by breeding true, and by causing linguistic isolation of population groups, it may be inimical to the Mendelian inheritance of strictly genetic traits.

Luckily, enough has been done in the

past few years to confirm the promise of these lines of investigation. There does appear to be a correlation between genetic traits and language, with occasional glaring exceptions.

Thus Papua New Guinea (PNG), with its 1,000 languages notoriously the world's citadel of linguistic diversity, appears to share with many primitive African populations a deletion of a tract of mtDNA not now found in Asian populations. Is that a sign that PNG was occupied by migration from Africa across the Indian Ocean, that the original migrants travelled by land but that the traces of their journey through Asia have been eliminated by later arrivals, or even that the deletion has been separately acquired by the two widely separated populations?

Whatever the truth, there seems little chance that such ambiguities will be resolved without more extensive investigations of the nuclear as well as mitochondrial genomes of the people concerned. The investigations so far involve such small numbers of people that they can be regarded only as pilot projects.

There is now an urgent need for fresh data, as well as for a project to collect and store DNA from small primitive populations still living in circumstances similar to those they would have occupied immediately after the great migration out of Africa. Already there is evidence that the genetic diversity of !Kung people of Africa (numbering a maximum of 10,000 people) is much greater than would be expected from the size of the population, suggesting a recent decline of numbers.

The prize offered by programmes of this kind is nothing less than a reconstruction of the cultural history of mankind. Is that not a benefit comparable in magnitude with some of the intended outcomes of the human genome projects? □

Gaps in molecular biology

FROM anatomy to zoology, there is no aspect of biology that has not been changed by the recognition that the essence of life lies in the DNA molecules in every cell, or in the nuclei of eukaryotic cells. Biology has become an open door at which an army of people is pushing energetically. Not much force is needed.

That is the general impression that this complex of research fields engenders in the public mind. Is the condition of ingrowing toenail hereditary? Never mind, it can only be a matter of time before somebody finds the gene, works out the nucleotide sequence and then offers gene therapy for the condition. And what is this talk of cerebral malaria in India? Are there not antibiotics now?

The outstanding consequence of the correct structure of DNA (Watson, J. D. & Crick, F. H. C. *Nature* **171**, 737–738; 1953) is not simply that it suggests a mechanism for inheritance, but that the structure is a means by which cells function as intended. At the outset, the second was an uncovenanted benefit.

The extraordinary achievements in biology in the past 40 years seem only to confirm the impression in the public mind that biology is on the down-slope of a helter-skelter. It is not surprising that it should be so. It is easy to underestimate the intellectual power that springs from understanding.

With all that said, a further and remarkable feature of the revolution in

biology has been that new techniques have been developed at a pace that has matched the need of them. Density-gradient ultra-centrifugation, the high-technology of 1960, is probably hardly mentioned to students any longer. But reverse transcriptase (1969), restriction enzymes (1970), the use of ligase (for the ligation of double-stranded DNA) and the polymerase chain reaction (PCR) in its many forms are their meat and drink. To keep an automatic gene sequencer serviceable requires that one should also know some computer technology.

Yet not everything has become child's play. On the contrary, there is a sense in which the questions buried under the glitz of recent success are the same as they have been for many years. What distinguishes one species from a close relative? How is the difference between one type of cell and another brought about? How (and why) did sexuality as a method of reproduction evolve, despite the associated metabolic costs? How does an embryonic bag of neurons become a brain, and is it a computer in the sense defined by Alan Turing?

What follows is not so much a celebration of four decades of quite remarkable success in research, but is meant to draw attention to some fields in which the molecular basis of life is not yet crystal clear. The underlying (and now common) assumption is that all life processes have a molecular explanation, at least in the

sense in which the weather has a mechanical explanation.

Another is that the research enterprise would benefit if molecular biology were more interdisciplinary than it has become, which is a strange criticism to make of a field in which many of the pioneers were ex-physicists (Perutz, Crick) and electrical engineers (Delbruck). *Nature* has for some time, but politely, complained that molecular biology is insufficiently quantitative, but there is more than that to say.

Both particular and general problems have been left aside in the scamper for information about the molecules involved in specific processes. The structure of DNA has been endlessly studied, but there appears not yet to have been an attempt to apply the now-sophisticated techniques of theoretical chemistry to the hydrogen-bonded purine-pyrimidine links that hold duplex DNA molecules together, for example. And while it is now more common than in the recent past for the importance of water interactions with protein (and other) molecules to be acknowledged, techniques for predicting their influence remain primitive.

More generally, there are gaps in the understanding of the energetics of life processes, all of which are sustained far from thermodynamic equilibrium by fluxes of sunlight and energetic chemicals. Luckily, the possible role of chaos in evolutionary processes has been recognized, while it now appears that several groups have independently built mathematical models of the cell cycle, whose ultimate importance will be great. □

The mechanics of life

THE laws of the mechanics of life are simplicity themselves.

(1) DNA and RNA are both made from four nucleotides strung together; DNA makes RNA makes protein (although sometimes RNA makes DNA). This was once known as the "central dogma".

(2) The genetic code is a triplet code; three nucleotides in DNA or RNA specify one amino acid in a protein.

(3) Genes are contiguous stretches of one of the two opposing strands of a duplex DNA molecule which may include meaningless stretches (which are dutifully removed from the transcribed messenger-RNA (mRNA) molecules by processing in the nuclei of eukaryotic cells).

As with newtonian mechanics, much of the interest of these statements resides in the uncertainties that persist. What follows is an incomplete listing.

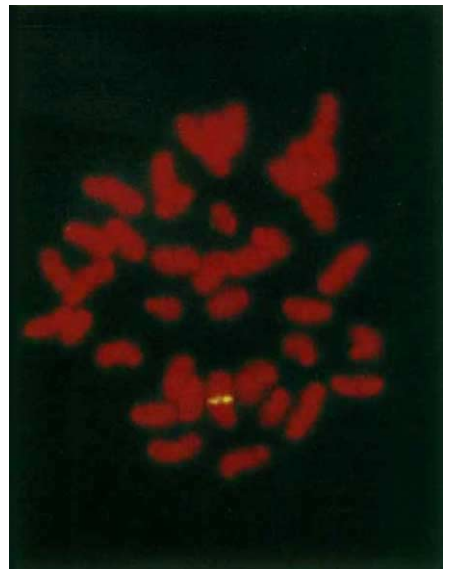
■ Proteins, the gene products, are assembled on (or perhaps 'in') cytoplasmic structures called ribosomes from the

information carried by mRNA. Despite three decades of investigation, there are only hypotheses to account for the functioning of ribosomes, themselves assembled from specific RNA and protein molecules.

■ Protein molecules naturally differ from each other in the sequence of their amino acids, but their function (as enzymes or otherwise) depends crucially on their shape. There is persisting uncertainty about the mechanism by which the molecules are folded into their active shape; for all but the smallest molecules, there are likely to be many inactive intermediate stages of folding.

One possibility is that the folding is partly determined by the order in which the amino acids are translated from the ribosomes, but there is also evidence that other protein molecules, called "chaperones", assist the process. The question is undecided, and there may be no general rule.

One bizarre twist to this problem is that



Human gene fragments (yellow) incorporated in the mouse genome and shown to be transmitted in the mouse germ-line (Jacobovits, A. *et al.* *Nature* **362**, 255–258; 1993).

the genetically abnormal protein responsible for the 'prion' diseases such as

Creutzfeldt-Jakob disease in people and bovine spongiform encephalopathy in the British cattle herd may function as a pathological chaperone molecule, inducing normal protein to precipitate in abnormal form.

■ Although the end-point for the transcription of a gene on a strand of DNA is usually well defined (by the presence of "stop" codons), its beginning is less obviously so. Transcription proper requires the assembly of a transcription complex, including the enzyme RNA

polymerase (a protein) and several other protein molecules, called transcription factors, some of which are required for the transcription of all genes, some of which are more specific.

While much is known of how the effectiveness of the transcription factors can vary with the metabolism of the cell, virtually nothing is known of the reasons why some genes in many cells are permanently inactive, as when their activity is not required in the organs to which the cells belong. □

The great leap forward

HAS the plan to compile the complete nucleotide sequence of the human genome turned biology into big science? That, some years ago, was the fear of many of those working in the field (not to mention those in other fields who feared that financial support for their work would be diminished).

That the expectation has proved correct is illustrated by the figure, showing a bank of automatic nucleotide sequencing machines, for all the world like the production line in some futuristic factory. Moreover, it seems likely that the small science grown big will become even bigger over the years ahead, as the world's pharmaceutical companies multiply each other's efforts.

But as yet there is no evidence that these developments have harmed the rest of biology, or of science generally. The chief development has been an increase of the scale of private spending on genetic research of all kinds. Time will tell whether the future will be different.

The ambition to produce a complete nucleotide sequence of all the human chromosomes goes back a quarter of a century, to the development (by Maxam and Gilbert on the one hand and Sanger on the other) of methods for sequencing short stretches of DNA. But when the small genomes of bacterial viruses were the most that people could hope to handle, producing a list of 3,000 million bases — the estimated number in the whole human genome — was plainly out of the question.

Curiously, there is still controversy about the number of genes contained in the 3,000 million nucleotides. The most commonly accepted estimate is that there may be 100,000 altogether, but many people argue that there are probably many fewer. Part of the uncertainty arises because, of the entire chromosomes so far sequenced (three from bakers' yeast, and part of the genome of the nematode worm *C. elegans*), between a third and a half

have no known function.

Whatever the number of human genes, there is no doubt that much of the DNA in the entire genome has no immediate functional purpose. At 1,500 nucleotides per gene, even 100,000 of them would occupy only 5 per cent of the total. But that is to neglect the saving of space in the genome by the use of the opposing strands of DNA as the repositories of genes and,



Nucleotide sequencing in action at the Institute of Genomic Research, in Bethesda, Maryland.

on the other hand, the likelihood that each gene will be preceded by a considerable region used by the transcriptional machinery.

Crick has famously called the large amount of DNA not involved in coding for the protein gene products "junk". Only the completion of the projects to sequence the whole genome will prove him right or wrong. But a part of the high promise of the complete sequence is that the junk conceals a record of the history of human evolution, perhaps represented by sequences derived from genes discarded in the course of evolution.

That the whole genome should be sequenced was part of the plan devised five years ago, both in the United States and by the private international organization called the Human Genome Organization (HUGO). At that stage, the molecular genetics community was fully aware of the difficulties ahead, and advocated an advance on a broad front — the

sequence of smaller genomes such as the chromosomes of yeast and of the plant *Aspergillus* and the nematode *C. elegans*.

Even that careful plan would not have been feasible without several earlier technical developments. The development a decade ago of artificial yeast chromosomes (YACs as they are called) makes it possible to include a stretch of DNA up to 1 million bases long in a synthetic construct that will multiply with yeast cells and from which identical DNA molecules can be recovered. This not merely a device for amplifying DNA, but for storing pre-defined DNA molecules with their dormant host cells. YACs have become the currency of inter-laboratory relations in genetics.

The development in 1983 of the polymerase chain reaction (or PCR) for amplifying shorter stretches of DNA, which has provided molecular genetics with a new confidence in the manipulation of fragments of DNA, has also provided the opportunity for the derivation of gene sequences from mRNA molecules taken from cells. This is the technique that has enabled J. Craig Venter and his organization, the Institute for Genomic Research in the United States, to produce short stretches of DNA, at most a few hundreds of nucleotides long, corresponding to the ends of genes.

These pieces of DNA (called expressed sequence tags, or ESTs) do not necessarily correspond to continuous stretches of DNA in the intact genome. Derived as they are from the mRNA molecules in cells, and called complementary or cDNA for that reason, they will tell nothing about the sequence of non-coding fragments that may be interspersed with the functional parts of genes.

Nevertheless it appears that this is the speediest way of making an index of the functional genes in the human genome. The snag is that it is merely an index, with the tags, or even only the number by which they are referred to, as the index entries. The full description of individual genes (if not already recorded in one of the data banks) must await detailed sequencing of the part of the genome concerned.

But there turn out to be benefits as well as disadvantages in the use of cDNA tags. Matsubara in Japan was among the first to use them to demonstrate the remarkable variation in the expression of genes from one tissue to another.

What has now emerged from Venter's work, covering several tens of tissue types (including embryonic tissue of different age) is that, in human beings, the brain is among the most prolific of sources of genes (or tags for genes) with no known function. That is a measure of how much remains to be found out about the mammalian nervous system.

There are other potential uses of this approach to genome sequencing, not least the opportunities for comparative anatomy. Is it too much to think that the cDNA approach to gene sequencing will be quickly embraced by those whose interest is in comparative evolution? It would, for example, be of great interest to know how many of the ESTs derived from human brain occur in the nervous tissue of the great apes.

But eventually there will be no substitute for the complete sequencing of the human genome. Part of the reason for the anxiety expressed by the academic community in the United States about recent developments stems from the fear that funds for the full project will dry up if the US Congress comes to believe that the sequencing project can be completed more quickly by other than the planned means, and with private support as well.

It goes without saying that the potential benefits for human health of an index to the human genome will be considerable. It is not so much a matter of recognizing genes that, when mutated, are responsible for disease, but of the prospects for using the genetic index as a guide to the understanding of the normal cellular physiology of the human body. But the scale of the effort that will be required for that enter-

prise will far exceed what the human genome project proper will require.

Much the same is true of the computational task that lies ahead. Already the management of the databases in which nucleotide sequences have been stored has taxed the capacity of the machines and those who attend to them. It is not merely the data stored, but that they must also carry annotations of where and how sequences of purportedly the same stretch of DNA have been obtained.

The computational problems are now enormously increased by the need to match long sequences of DNA (represented by tags, for example) against stretches of the genome approaching a few per cent of its total length. The resources that will be required to manage the genomes of other mammals than people are not easily anticipated.

That is yet another reason why the case for the whole project is not diminished by the prospect that a complete index to the genome will soon be available. Indeed, to the extent that mutations of the non-coding sections of human DNA may affect the fidelity, or the ease, with which the genome as a whole functions in certain kinds of cell, it is entirely possible that the causes of many human conditions will not be identified until the whole job is done. □

tions. By general consent, the representation in 1975 by the late David Marr (who died in 1980 at too early an age) of the working of the human visual system as a problem in artificial intelligence has helped refine the questions people ask of how the visual system makes use of, or processes, information.

Still more remarkably, the earlier suggestion by Professor John Hopfield of the California Institute of Technology that the functioning of the brain, and especially of the cortex, could be represented by the behaviour of networks of electrical elements interacting with different strengths turns out to be of practical value.

The inspiration of the idea is the observation that the processing power of the brain appears to be widely distributed. The interactions, of various strengths, between neighbouring elements in the network represent the varying ways, and no doubt degrees, in which synaptic connections are established between neighbouring neurons.

The reason why there is no general agreement on how (or even whether) the brain can faithfully be represented by any particular model of this kind is almost certainly that the question is ill-posed. It is a matter of resolution. While electrical recordings from single cells are possible, present techniques will not suffice to record from whole groups of neighbouring cells so as to identify the functions they perform in concert, and the other functions that they share with members of other groups.

Yet neural networks have found a life of their own by being transferred to chips of silicon which are designed to perform tasks resembling some of those done in the brain.

Is the brain a computer?

To a crude approximation, present understanding of how the brain is formed in the course of development and of how it functions in the adult is almost non-existent. Several decades of hard and productive work in the neurosciences have uncovered much about the anatomy of the mammalian (and even the human) brain, but functional understanding awaits conceptual definition.

That strong statement should not be taken to decry the achievements of recent decades. People such as Hubel and Wiesel, in the 1960s, did for neuroscience what Hubble did for cosmology three decades earlier by demonstrating the intricacy and complexity of the problem of the central nervous system. That neurons are ramified structures, extending over millimetres, was a revelation comparable in its effect with the recognition that the Universe is expanding.

Electrical recording from individual cells has also revealed the complexity of the connections between sense organs and the various regions of the cortex. That the intact brain incorporates a large degree of redundancy is evident from the many observations of how people whose brains have been damaged nevertheless retain much of their function.

But why the mammalian cortex should, for example, include several regions to which the visual system makes projections of the outside world in colour is a further complication whose function is not clear. The standard explanation is that the several regions are interrogated (to use a computer word) in different ways by regions of the brain needing access to the information stored in them.

The computer analogy is almost irresistible, as it has been since J. von Neumann's first essays in the design of electronic computers in the early 1940s. In the broad sense, the assertion is undeniable. The system takes in information through the senses which, as electronic input-output devices go, are relatively slow, operating on a timescale of milliseconds, but the use made of the information is equivalent to that of a calculation.

Using the head, a person can avoid an obstacle in his or her path, can make a decision among alternative ways of getting most conveniently from one point to another, and can even perform mental arithmetic — literal calculations.

So it is natural that there should have been several attempts to represent brain functions symbolically as computer func-

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

False-colour scanning electron micrograph of neurons from the cerebral cortex. Magnification $\times 570$.

None of this implies that there is no purpose in pursuing the analogy between the brain and a computer of a kind, but

simply that too little is known of how the brain is constructed, and how it becomes what it is, for the analogy to be fruitful.

Indeed, the attempts that have been made to build models of single neurons, so as to represent their electrical response in terms of the external influences to which they are subjected, yield only poor results. There are simply too many channels of influence, electrical and otherwise, on the surface of a neuron, for its behaviour to be modelled in other than a symbolic fashion in the present state of observation.

The question remains of how the human brain becomes what it is. Observations with newborn mammals have shown that even the visual system remains plastic for several weeks after birth, presumably because neuronal axons have to learn to find their way to appropriate partners. That, fortunately, is a process that can be studied directly.

For people, the practical question will be to form a general view of how long the various functions of the brain remain susceptible to learned influences. Among other things, that may help to set psychology, for too long a Cinderella science, on a firm foundation for the first time in history. □

Promise of biodiversity

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Rainforest in Queensland, Australia, typical of the terrain protected by the Biodiversity Convention signed at Rio de Janeiro in August 1992 (but already safeguarded by the Australian federal government). Devising ways of estimating biodiversity quantitatively remains an unsolved problem.

Penny Tweedie/ Panos Pictures

Energy source for cells



Electron micrograph of a single mitochondrion from a human pancreas cell, showing the cristae (internal folds of the inner membrane) at which respiration occurs. Also visible is the endoplasmic reticulum within the mitochondrion, carrying bead-like ribosomes (for protein synthesis). Mitochondria have their own complement of DNA, responsible for the production of some mitochondrial proteins, which replicates independently of the nuclear DNA. Mitochondria are believed to have derived from Archaean symbiotic bacteria. (Science Photo Library)

Molecules in embryology

THERE is a well-known gene in the fruit-fly *Drosophila* called *eyeless*, whose mutants have no eyes. In the mouse, there is a similar ('homologous') gene called *small eye*, whose mutants have small eyes. Yet the fly's eye is a compound eye and the mouse eye an integral structure like that of other vertebrates.

So much was reported at the annual meeting of drosophilists at Crete this year by Walter Gehring, of the Biozentrum at Basel, the one who discovered more than a decade ago the genetic hallmark of the class of genes known as homeotic genes — a short sequence of DNA upstream of the gene proper and now known as the "homeobox".

This unexpected finding is in itself a curiosity. But it is a vivid sign that the development from embryo to adult of two very different organisms may have something substantial in common with each other. If nothing else, this justifies the

long-standing use of *Drosophila* as the work-horse of the study of development.

In reality, two kinds of study are intertwined. Development is another name for embryology, which includes the metamorphoses that insects undergo. But embryonic development leads to the specialization of cell types, and thus to the silencing of inappropriate genes (by a mechanism not yet understood). There are two switches to find: on and off.

Molecular techniques have enormously helped to transform both fields in the past few years. The search now is for the diffusion of gene products of vertebrate homeotic genes that play parts analogous to those of the gene products of the corresponding genes in *Drosophila*, which appear by their movement through the body of the developing embryo to regulate the specialization of cells. This is a door that will need little pushing to swing open. □

In praise of the scientist as writer

Prizes of £500 are awarded to Dr Paul Grant and Dr Dan Hultmark, each of whom has provided *Nature* with an outstanding News and Views article during its 125th anniversary year.

ON accepting an invitation to write a News and Views article for this journal, or on proposing one, authors are sent a single sheet of paper dense with advice as to how to set about the job. These guidelines point out that the News and Views section provides a forum for the dissemination of scientific news to a wide audience. They go on to say that authors should make clear the nature of the advance concerned, which is usually described in a paper in *Nature* or elsewhere; and that, while communicating a sense of excitement, they should also provide a critical evaluation of the work.

Often, the guidelines are accompanied by a no-doubt irritating letter, from one or other of the editors who work on the section, reiterating the cardinal principle — that the news is of the essence, and that the way to tempt the uncommitted reader to read on is not to set the stage with background information, as maybe in an essay or review, but in a journalistic manner to put the nub of the research at the beginning. This approach rather cuts against the grain of most professional writing. Nonetheless, it generally works. Anybody wishing to judge the results can do so by browsing through some back issues.

Why, then, should one of the two prizes for literate and interesting News and Views articles published during *Nature*'s 125th anniversary year — 4 November 1993 to 3 November 1994 — go to a writer who blithely ignored that advice? He began:

In December 1986 I boarded a plane in San José, California, bound for Zürich, determined to unearth the facts behind the reports and rumoured confirmations of superconductivity at 30 K which had recently been published by Georg Bednorz and Alex Müller.

Eighteen hours later, I stumbled jet-lagged into the IBM Ruschlikon laboratory. There I found an agitated and preoccupied young man — Bednorz had just received the preprint of Kitazawa and Tanaka's confirmation, and, justifiably, was concerned that the rest of the world would run away with his discovery.

Within half an hour, my travel weariness had disappeared. The amount and quality of the data my Swiss colleagues had accumulated was far beyond that generally known at the time, leaving no doubt that superconductors with T_c as high as 40 K really existed. I returned home a true believer, with the peace of my holiday delightfully shattered, ready to spread the good news to my colleagues in IBM Almaden.

This was Paul Grant, now at the Electric Power Research Institute, Palo Alto, California, writing in the 6 January 1994 issue.

The prompt for the anecdote was the publication, seven Decembers on from 1986, of two papers with claims for dramatic new highs in the transition temperature (T_c) of superconducting materials.

Grant earns his prize not because he tore up the rule book with such panache. Rather, in a single printed page, including references, he went on to examine the claims in a splendidly readable yet authoritative manner. That he had reservations about them was a demonstration of the demands on an author of secondary scientific writing such as this to point to possible inconsistencies and flaws in research papers, and to alternative interpretations to those put forward by the researchers concerned — painful though it may be to them.

The article, "Another December revolution?", is beautifully clear, for all that in the body of the piece Grant is necessarily involved in technical detail; and the narrative impetus stemming from the opening paragraphs picks up again towards the end with a judicious conclusion (which was wise, given the forceful form of the rest of the text).

Above all, the article is written to be read and enjoyed, as well as to convey information; it is not a referee's report. Although the thinking is rigorous, the prose is relaxed. 'Scientese' is a stiff language for the formal paper and lecture theatre, and in communication for non-specialists (but not, in News and Views, for non-scientists) it fails to make up in precision what it loses in simplicity or richness of expression. Putting the words 'scientific' and 'literature' together is not often justified. In the case of Grant's article it is.

It is not so long since digestible, independent and expert comment on science emerged into print in periodicals. In *Nature*, News and Views started to take the form it now has only in the late 1960s, most of the contributions then coming from regular (and anonymous) correspondents or staff members. Since then anonymity has gone and colour graphics have arrived. Regular correspondents are fewer than they were, and publishable unsolicited articles — which remain welcome — are comparatively rare. This may be a sign of the times. As the pace of research and competition for financial support has intensified, the need for scientists to write papers and grant proposals, rather than think up ideas for writing disinterested articles about the work of others, has become all the greater.

All of which is to say that whereas the short list for the prize won by Paul Grant, that for a solicited News and Views article, was in fact a long list, there were fewer contenders to choose among for the prize for an unsolicited contribution. Nonetheless the winning piece is every bit as worthy as "Another December revolution?". As in that article, the subjects for comment were papers published in journals other than *Nature*.

"Ancient relationships", by Dan Hultmark of Stockholm University, appeared under the rubric "Insect immunology" in *Nature* of 13 January 1994 (oddly enough, just a week after Grant's article). There is no need to paraphrase the gist of his story, for he did it himself in straightforward style:

Insects look nothing like vertebrates and their organ systems seem to be built on entirely different principles. Nevertheless, as we get a better understanding of how these systems operate at the molecular level, unexpected similarities are emerging. Among them must now be counted similarities in the respective immune defences, as reported in two recent papers.

So readers know where they are about to go. The article's other strengths are that the prose is plain, the explanations are clear without seeming didactic, and the text is succinct (again, no more than a printed page). Hultmark himself produced a helpful diagram which summarized the main points in visual form.

Here, though, it is the structure of the article that stands out. The opening is followed by an account of the underlying research agenda and of antecedent work, before Hultmark gets to grips with the detail of the papers concerned. Then comes a classic (and crucial) passage dealing with questions thrown up by the new developments, and finally a closing flourish. As in fiction, the best News and Views articles have an identifiable beginning, middle and end. "Ancient relationships" is as good an example of that as any piece, solicited or unsolicited, published over the past 12 months.

Choosing these two winners has been a thoroughly enjoyable but invidious task. Since 4 November 1993, some 350 other busy people have been in receipt of the guidelines for authors and have found the time to write for News and Views — for that they are all owed *Nature*'s earnest thanks.

Tim Lincoln

Tim Lincoln is News and Views Editor of Nature.

Unveiling a new galaxy

David Burstein

ON a clear, dark, moonless summer (or winter) night, look at the sky around midnight. You will see myriads of stars pasted on the dome of the sky, bisected by a fuzzy arc that is an amalgam of dark and light diffuse patches. This arc was called the 'Milky Way' by the Greeks. We know today that the Milky Way is the disk of our Galaxy, seen edge-on by us, residents of that disk. Yet what is beautiful to the eye on a dark night is the bane of extragalactic astronomers, as the dust and stars within the disk of our Galaxy veil the rest of the Universe from our sight. Extragalactic astronomers have long called the Milky Way the 'zone of avoidance' for the fact that galaxies appear to avoid that region of the sky. This is an illusion, of course. Only this summer, we first realised that the Milky Way had an attendant dwarf galaxy so close that it was in danger of being pulled apart¹; and now, on page 77 of this issue, R. C. Kraan-Korteweg *et al.*² describe their discovery of a nearby spiral galaxy hidden by the murk.

If we are truly going to understand the structure of the Universe, we must be able to study the 15 per cent or so that is obscured by our own Milky Way. The trick is to search the region at a wavelength that is not badly obscured by dust, and the Dwingeloo telescope in The Netherlands is, for the time being, entirely dedicated to doing just that. A systematic survey of the zone of avoidance (ZOA) region is being made at a wavelength of 21 cm (the emission comes from atomic hydrogen, H I). Analysis of the results is only just beginning, and the first fruits of the survey include the new spiral galaxy. More can surely be expected to emerge as the survey continues.

The new galaxy, seen glimmering faintly behind the foreground stars in the photograph, lies about 0.1° below the Galactic plane and has an angular size of about 3 arcminutes as imaged by the discoverers. As far as Kraan-Korteweg *et al.* can guess, it is about 3 megaparsec away (about 10⁷ light years), too distant to be a member of our Local Group of galaxies. It therefore is not likely to share the fate of the dwarf galaxy discovered earlier this year, which is a mere 24 kiloparsec away and is apparently being pulled apart by the Milky Way.

It is indicative of the importance placed on understanding the Universe beyond

the ZOA that the Dwingeloo survey is but one of several being done by astronomers, as a meeting earlier this year³ discussed. Surveys in the optical, infrared and radio regions of the spectrum are being done in various directions within the ZOA by various groups. Indeed, the paper on page 77 draws together workers from several groups; Lahav and Lynden-Bell have had



A galaxy on our doorstep — image of Dwingeloo 1 taken by the Isaac Newton Telescope.

a long-standing cosmological interest in what lies behind the ZOA, Henning was earlier involved in a more limited H I survey aimed at detecting nearby galaxies, and to the best of my knowledge, Kraan-Korteweg herself was the first to start systematically examining optical images for galaxies within the ZOA, some six or seven years ago.

Why this flurry of interest? After all, astronomers have known about the obscuration of the disk of the Milky Way for nearly a century. Speaking for myself, my attention was first drawn to the ZOA seven years ago, when it became apparent that one of the largest components of the motion of our Galaxy, relative to the expansion of the Universe, is due to the gravitational attraction of mass that lies behind the ZOA. Moreover, the largest nearby concentration of visible mass that we do know exists (called variously the Great Attractor or the Hydra-Centaurus-Pavo-Indus-Telescopium collection of

galaxy clusters) is squarely bisected by the ZOA, leaving us to speculate on how many galaxies are really in this region.

Whereas the disk of the Milky Way is the zone of avoidance to those who study other galaxies, it is a subject in its own right to those who study our Galaxy. Over the past 10 years, the advent of near-infrared and infrared detectors with high quantum efficiency that can pierce the dust have made observational study of the disk plausible. Inevitably, such studies are bound to uncover extragalactic objects lying behind the ZOA; for example, Ibata and colleagues¹ discovered the Sagittarius dwarf galaxy during a survey that was designed to study stars within the disk and bulge of our Galaxy.

Study of the stars and dust in the disk of our own Galaxy is important for other cosmological reasons as well. The discovery of the faint lumps and bumps of the early Universe by the COBE satellite first required careful removal of the much brighter foreground infrared emission from both stars and dust in the disk of our Galaxy (the newly discovered galaxies, on the other hand, are far too faint to affect the COBE results). COBE, indeed, contributes considerably to our knowledge of the Galactic disk⁴. Separately, large-scale surveys that are designed to find the dark-matter candidates known as MACHOs (massive compact halo objects; really, very low-mass star-like objects) must observe as many stars as possible in order to have a reasonable probability of detecting the light variation caused by passage of one MACHO along our line of sight to one star. So

MACHO surveys place high priority on relatively dust-free paths within the disk and bulge of our Galaxy.

Until 10 years ago, most maps of galaxies on the sky placed a veritable 'celestial incognito' on the ZOA; this region was simply kept blank. Ten years ago our understanding of the distribution of stars in the disk more than 2 kpc from us was limited to a very few lines of sight and to studies of only the brightest stars. It is encouraging to see the veil of the Milky Way being slowly lifted; and we await with anticipation the full unveiling. □

David Burstein is in the Department of Physics and Astronomy, Arizona State University, Tempe, Arizona 85287, USA.

1. Ibata, R. A., Gilmore, G. & Irwin, M. J. *Nature* **370**, 194–196 (1994).

2. Kraan-Korteweg, R. C. *et al.* *Nature* **372**, 77–79 (1994).

3. *Unveiling Large-Scale Structures behind the Milky Way* (eds Balkowski, C. & Kraan-Korteweg, R. C.) (Astr. Soc. Pacif. Conf. Series **67**, 1994).

4. Arendt, R. *et al.* *Astrophys. J.* **425**, L85–88 (1994).

Hammerhead nailed down

Thomas R. Cech and Olke C. Uhlenbeck

EVER since 1982, when RNA was first found to catalyse the specific cleavage and formation of chemical bonds, we have wanted a detailed picture of such an RNA — a ribozyme. Protein crystallographers have provided numerous examples of enzyme active sites built from amino acids. But we have had no similarly detailed structures for catalytic centres constructed from nucleotide building blocks. The X-ray structure of the hammerhead ribozyme, described by David McKay's group on page 68 of this issue¹ will thus be enthusiastically received by those interested in macromolecular catalysis as well as those working to understand RNA structure.

The hammerhead occurs embedded in various RNAs that infect plants², where its self-cleavage activity is thought to process replicative intermediates. This structural motif consists of three base-paired stems emanating from a central core of conserved sequence, superficially resembling the handle and head of a carpenter's hammer³. Cleavage of a phosphodiester bond occurs at a unique site in the motif, generating 2',3'-cyclic phosphate and 5'-hydroxyl termini. A multiple-turnover catalyst can be produced by separating the substrate strand (the one that undergoes cleavage) from the ribozyme (the catalytic entity) (Fig. 1). In the version of the hammerhead that was crystallized¹, the substrate RNA strand was replaced by its DNA analogue. DNA is not cleaved owing to the absence of the 2'-hydroxyl attacking group at the cleavage site.

In three dimensions, the ribozyme is not so much a hammerhead as a wishbone¹. Stems I and II diverge from the core at an acute angle, while stem III points in the opposite direction. The central core is sewn together by base-pairing interactions that are neither Watson-Crick nor wobble pairs (Fig. 2). Although there are now a number of RNAs in which 'internal loops' have been found to be highly structured⁴⁻⁷, there are still no rules for predicting the actual fold. Structural biology is necessary.

One domain of the core, consisting of the conserved CUGA sequence, is involved in the sharp turn of the ribozyme strand at the base of stem I. Remarkably, it is virtually identical to the U-turn seen in the anticodon and T loops of transfer RNA. The other domain, consisting of conserved nucleotides at the base of stem II, forms a basket for metal-ion coordination in its widened major groove. Its key structural element consists of consecutive G·A pairs⁸ with the top edge of each A paired with the bottom edge of a G,

similar to a structure seen in a model duplex⁹. When Mn^{2+} or Cd^{2+} was soaked into the crystal, the metal ion was observed to be suspended between the N-7 of a guanine base and a phosphate in this domain (Fig. 2b). This metal ion is distant and seemingly inaccessible to the cleavage site in the structure, so it presumably has a stabilizing rather than a catalytic function.

escence energy transfer¹⁰. The derived orientations of the helices agree well with the crystal structure. In the other, different pairs of the three helices were elongated by about 70 base pairs, and the angle between each pair was estimated by transient electric birefringence¹¹. Although the helix angles obtained do not agree as well with those seen in the crystal structure, this experiment was unique in that a cleavable RNA substrate strand was used. Further confirmation of the hammerhead structure is eagerly awaited, but meantime we note that at least one natural hammerhead has a pseudoknot connect-

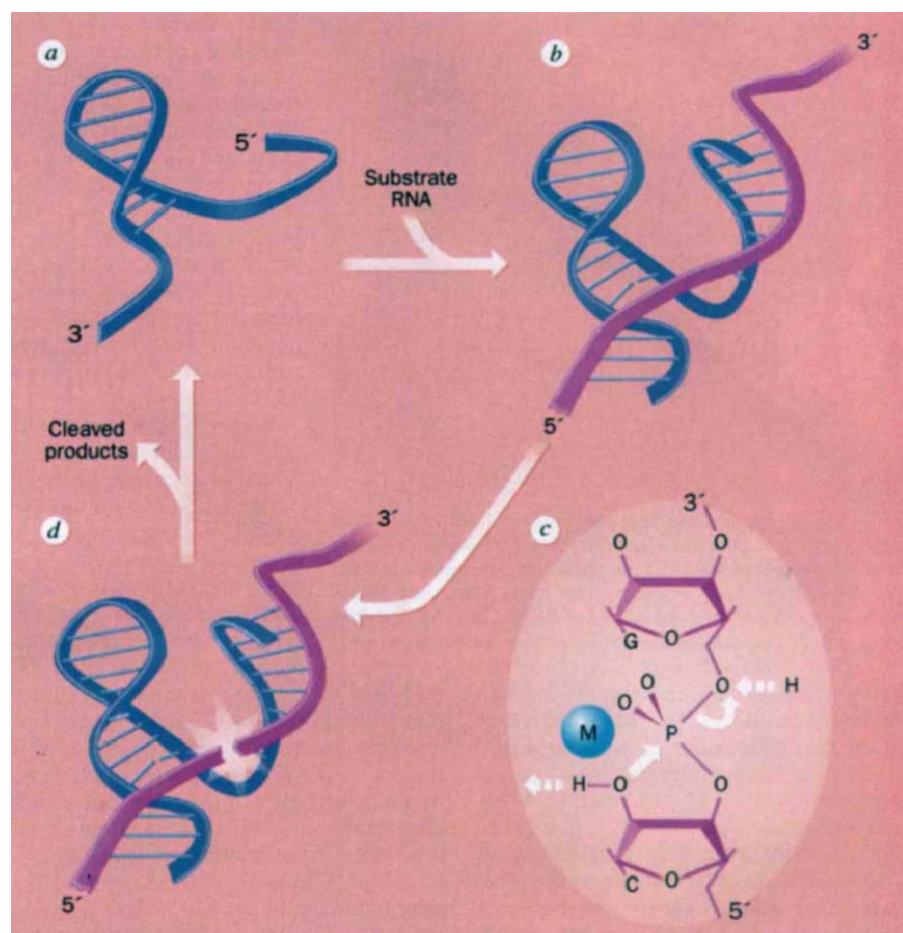


FIG. 1 The hammerhead ribozyme catalytic cycle can now be shown in three dimensions. The free ribozyme (a) binds its substrate RNA (pink) by base-pairing (b). Cleavage occurs by transesterification (c) assisted by a divalent metal ion (M), resulting in the ribozyme-product complex (d). Product release completes the cycle.

As with any crystal structure, one must consider the possibility that crystal-packing forces could distort the molecule from its solution conformation. This is less of a concern than usual for the hammerhead, because each unit cell contains three very similar molecules that are subject to different crystal-packing constraints. Two different solution experiments to assess the relative orientation of the three hammerhead helices have been carried out. In one, different pairs of helices were equipped with fluorescent groups and their relative distances measured by fluor-

ing helices I and II (ref. 12), consistent with their proximity as observed in the crystal structure.

The crystal structure neatly explains many of the biochemical experiments defining the structural requirements for hammerhead activity. Innumerable mutagenesis experiments have established that any base pair will serve in most of the helical positions, but few of the core residues can be changed without reducing the cleavage rate¹³. In most cases it is clear from the structure that such mutations would disrupt the complex network of

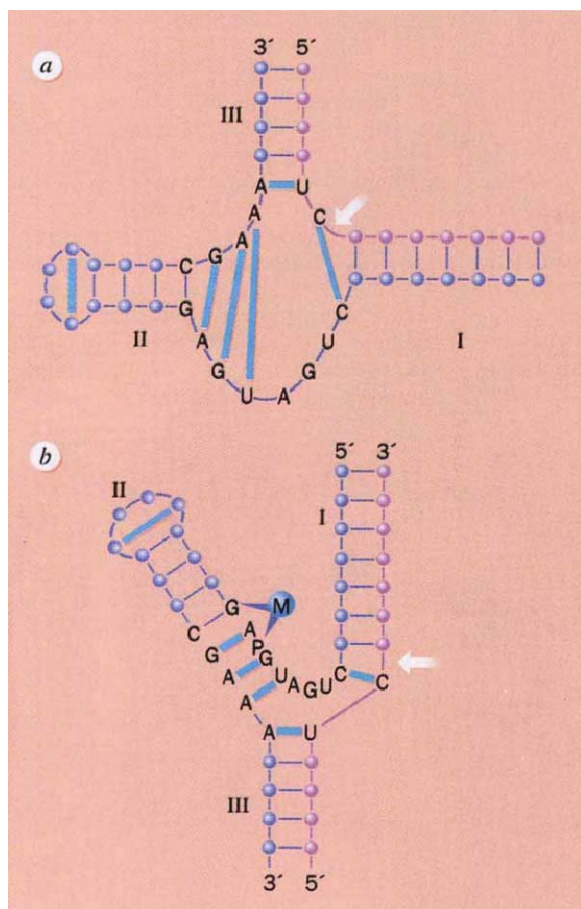


FIG. 2 The hammerhead ribozyme (blue) paired with its RNA substrate (pink) in the old secondary structure model (a) and a schematic representation of the structure determined by Pley *et al.*¹ (b). Thin lines, Watson-Crick base pairs that define stems I, II and III; thick bars, novel base pairs; M, divalent metal ion; arrow, site of ribozyme-catalysed cleavage.

hydrogen bonds in the core. A similar conclusion can be reached when analysing the substantial number of more sophisticated experiments where specific atoms or functional groups in the hammerhead were altered¹⁴. The relatively few cases where the biochemical data appear to be at odds with the crystal structure may be the most interesting. Of course, the reaction rate monitors the structure of the transition state of a reaction, whereas any crystal structure pertains to a ground-state complex. As an alternative explanation, the use of the non-cleavable substrate analogue may produce a ground-state structure slightly different from that which leads to cleavage.

As is often the case when the structure of a protein enzyme is first solved, the crystal structure of the hammerhead has not immediately revealed the mechanism of catalysis. Like the protein ribonucleases that generate 2',3'-cyclic phosphates, hammerhead cleavage proceeds with inversion of stereochemical configuration about the phosphorus atom. This suggests that an 'in line' reaction occurs through a transition state (or short-lived intermediate) in which the attacking

2' oxygen is on one apex of a trigonal bipyramid and the leaving 5' oxygen is on the opposite apex (Fig. 1c). In the crystal structure, the orientation of the 2' and 5' oxygens is very different from that needed for in-line attack. Pley *et al.*¹ speculate that the problem may in part be due to the absence of the attacking 2' oxygen and show that a relatively modest reorientation of the structure would provide a configuration that could react. There is also biochemical evidence for a critical divalent metal ion bound to one of the oxygens of the scissile phosphate¹⁵. Although the resolution of the crystal structure is too low to see a Mg^{2+} ion, it is puzzling that no Mn^{2+} ion was seen at this site. Clearly there is more for crystallographers and biochemists to do before we understand how the hammerhead works.

Bigger RNAs are able to solve a problem not encountered in the hammerhead motif: sites distant in the secondary structure engage in specific interactions. One often-used solution involves hairpin loops of sequence GNRA (where N is any nucleotide, and R is A or G;

ref. 16). These loops have a defined structure of their own¹⁷, and they can also dock into minor grooves of distant double helices¹⁸. The X-ray structure of the hammerhead fortuitously illuminates just such an interaction — as McKay's group shows in another paper in this issue (page 111)¹⁹, the GAAA loop of stem II contacts the minor groove of stem I of a different molecule in the crystal lattice¹⁹. The sequence covariations originally observed by Michel and Westhof are explicable by the

new structure, although the pattern of hydrogen bonding differs from the base triples originally modelled¹⁸. One striking feature of the tertiary interaction is the extensive network of hydrogen bonds between the bases and ribose 2'-hydroxyl groups¹⁹. It seems entirely reasonable that 2'-hydroxyls should serve as important 'handles' for RNA recognition²⁰, because once a nucleic acid helix forms, the most accessible hydrogen-bonding moieties are the phosphates and (in RNA) the 2'-OH groups.

Because of its small size and the ease with which it can be adapted to cleave any target RNA sequence²¹, the hammerhead has been the favourite of those developing ribozymes for therapeutic purposes. The goal is to synthesize modified hammerheads that are metabolically stable and can enter cells to destroy a 'target' viral RNA or messenger RNA involved in promotion of a disease. The X-ray structure will lead to more sophisticated ribozyme engineering; for example, the G·A pairs could be replaced by an analogue that might alter the divalent cation requirement, or even stabilize the structure in the absence of metal ions. So, like rational drug design, rational ribozyme design should now be possible.

Twenty-one years ago the structure of yeast tRNA^{Phe} was solved. It remains one of the most influential events in RNA science. Since then, new RNA structures have been limited to other tRNAs and several small RNA motifs. Just as protein crystallography accelerated when recombinant DNA and high-level expression systems became available, RNA crystallography is now being stimulated²² by the development of *in vitro* transcription and solid-phase chemical synthesis methodologies. In all, the hammerhead structure is a harbinger of a new era in the structural biology of RNA. □

Thomas R. Cech and Olke C. Uhlenbeck are in the Department of Chemistry and Biochemistry, Howard Hughes Medical Institute, University of Colorado, Boulder, Colorado 80309, USA.

- Pley, H. W., Flaherty, K. M. & McKay, D. B. *Nature* **372**, 68–74 (1994).
- Prody, G. A., Bakos, J. T., Buzayan, J. M., Schneider, I. R. & Bruening, G. *Science* **231**, 1577–1580 (1986).
- Forster, A. C. & Symons, R. H. *Cell* **49**, 211–220 (1987).
- Szewczak, A. A., Moore, P. B., Chan, Y. L. & Wool, I. G. *Proc. natn. Acad. Sci. U.S.A.* **90**, 9581–9585 (1993).
- Wimberly, B., Varani, G. & Tinoco, I. Jr *Biochemistry* **32**, 1078–1087 (1993).
- Battiste, J. L., Tan, R., Frankel, A. D. & Williamson, J. A. *Biochemistry* **33**, 2741–2747 (1994).
- Peterson, R. D., Bartel, D. P., Szostak, J. W., Horvath, S. J. & Feigon, J. *Biochemistry* **33**, 5357–5366 (1994).
- Gautheret, D., Konings, D. & Gutell, R. R. *J. molec. Biol.* **242**, 1–8 (1994).
- Santa-Lucia, J. J. & Turner, D. H. *Biochemistry* **32**, 12612–12623 (1993).
- Tuschl, T., Gohlke, C., Jovin, T. M., Westhof, E. & Eckstein, F. *Science* **266**, 785–789 (1994).
- Amiri, K. M. A. & Hagerman, P. J. *Biochemistry* (in the press).
- Miller, W. A. & Silver, S. L. *Nucleic Acids Res.* **19**, 5313–5320 (1991).
- Ruffner, D. E., Stormo, G. D. & Uhlenbeck, O. C. *Biochemistry* **29**, 10695–10702 (1990).
- Bratty, J., Chartrand, P., Ferbeyre, G. & Cedergren, R. *Biochim. biophys. Acta* **1216**, 345–359 (1993).
- Dahn, S. C. & Uhlenbeck, O. C. *Biochemistry* **30**, 9464–9469 (1991).
- Woese, C. R., Winker, S. & Gutell, R. R. *Proc. natn. Acad. Sci. U.S.A.* **87**, 8467–8471 (1990).
- Heus, H. A. & Pardi, A. *Science* **253**, 191–194 (1991).
- Michel, F. & Westhof, E. *J. molec. Biol.* **216**, 585–610 (1990).
- Pley, H. W., Flaherty, K. M. & McKay, D. B. *Nature* **372**, 111–113 (1994).
- Pyle, A. M., Murphy, F. L. & Cech, T. R. *Nature* **358**, 123–128 (1992).
- Haseloff, J. & Gerlach, W. L. *Nature* **334**, 585–591 (1988).
- Doudna, J. A., Grosshans, C., Gooding, A. & Kundrot, C. E. *Proc. natn. Acad. Sci. U.S.A.* **90**, 7829–7833 (1993).

Conveying past climates

Ed Boyle and Andrew Weaver

THE Atlantic Ocean's great 'conveyor belt' circulation, which carries warm surface water north, and cold, dense 'deep water' south, has been viewed as a possible source of rapid climate shifts. Could the conveyor have been 'switched off' in the past by injections of fresh water from melting ice caps? Numerical ocean models have tended to suggest so, producing a range of behavioural modes with very different rates of deep-water formation in the North Atlantic. But on page 82 of this issue, Stefan Rahmstorf¹ shows that a seemingly small change to the treatment of heat flux in one such model permits it to settle into any of three distinct modes of operation with the same deep-water formation rate.

Once a stable North Atlantic circulation resembling the modern ocean had 'spun up' in the model, very brief (4-year) pulses of fresh water at high latitudes briefly diminished overall convection. After the freshwater forcing was stopped, however, convection was quickly restored with the twist that in the new stable state, the North Atlantic convection cell was shallower with, below it, a stronger deep inflow of Antarctic bottom waters (see figure). The North Atlantic climate that emerged was considerably colder than today's. Following the jump into this 'shallow conveyor' mode, another brief freshwater pulse induced a third mode of stable flow somewhere between the first two modes.

The idea that the present-day ocean circulation may not be the sole stable state goes back many years. Chamberlin² noted the delicate balance between thermal and haline effects in driving the present-day conveyor and speculated that different modes of operation may have occurred in the past. Many years later, Stommel³ quantitatively showed that a convective fluid regime (with independent temperature and salinity density control) could have multiple stable modes of circulation under constant atmospheric forcing. Other multiple-mode fluid regimes were seen in later model studies (see review in ref. 4). More recently, Broecker *et al.*⁵ proposed that rapid climate fluctuations observed in the Greenland ice-core records might be due to abrupt 'on' and 'off' changes in the flow of the North Atlantic Deep Water

(NADW) circulation. Soon thereafter, Frank Bryan⁶ discovered that the ocean general circulation model code developed by the Geophysics Fluid Dynamics Laboratory (GFDL), when applied to a simplified flat-bottomed single ocean basin extending from pole to pole, showed two stable states. One of these states had convection at both poles, the other had convection at one pole only. Manabe and Stouffer⁷ used the GFDL coupled ocean/

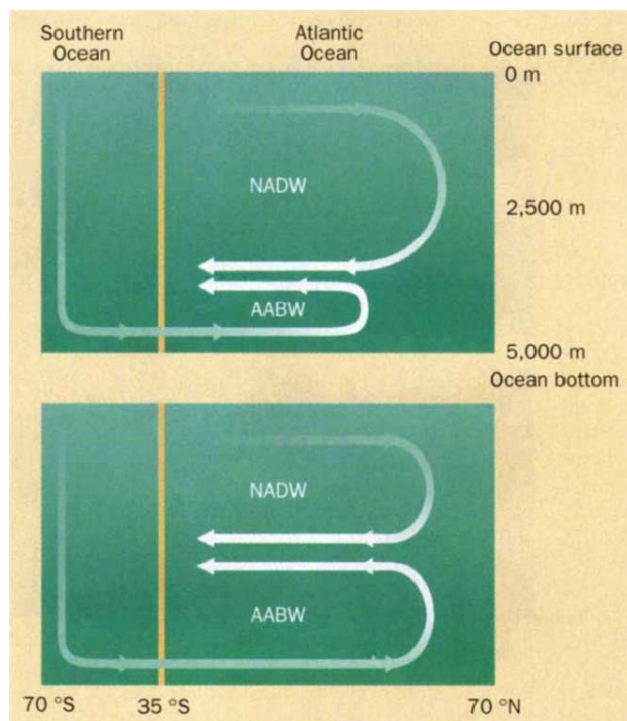
correction and instead forcing towards some predetermined surface ocean temperature)⁴. This new model instead calculates a heat flux between the ocean and atmosphere in two terms: first, a radiative term that is sensitive to the local temperature difference between the atmosphere and sea surface; second, a term depending upon the laplacian of the temperature difference, which aims to account for the 'diffusive' transport of heat away from anomalously warm surface waters.

The new parameterization should be viewed as a significant step in the right direction. As pointed out by Bretherton⁸, latent and sensible heat losses are inefficient mechanisms for the removal of large-scale anomalies in sea surface temperature; heat lost over one part of the ocean must be advected by the atmospheric winds away from the anomaly for it to be effectively removed. Rahmstorf's new formula incorporates these ideas in such a way that in the limit of a worldwide anomaly in ocean surface temperature, radiative heat loss is the only mechanism for its removal.

Initially one might expect that in the real world, convection and NADW formation would be strongly coupled so that it would be impossible for convection to change without corresponding changes in the NADW formation rate. This idea that convection acts as a conduit through which surface waters are converted to bottom waters has recently been shown to be false. Instead, Send and Marshall⁹ have shown that convection mainly acts to mix temperature and salinity vertically. This vertical mixing sets up basin-scale horizontal pressure

gradients in the ocean which then drive the ocean conveyor. This work justifies the convective parameterizations currently used in coarse-resolution ocean models and gives us faith that Rahmstorf's results provide a plausible picture of past ocean circulations.

From the perspective of a palaeo-oceanographer, an interesting feature of this new model is that its shallower circulation mode resembles that seen in reconstructions of oceanic surface temperature and salinity^{10,11} from foraminiferal species variability, and oxygen isotope measurements and deep chemical distributions inferred from the composition of fossilized deep-sea organisms¹². In the modern North Atlantic, biologically active nutrients such as phosphorus, nitrogen and silicon are scarce because of the flushing



Schematic diagram of North Atlantic Deep Water overlying deeper Antarctic Bottom Water. Top, present-day circulation; bottom, possible circulation in the Last Glacial Maximum, with a shallower North Atlantic conveyor and stronger flow of Antarctic Bottom Water.

atmosphere model with realistic geography and topography, and found that with the same flux conditions at the surface, NADW was stable in both 'on' and 'off' modes. Since that time, numerous other model studies have uncovered other multiple modes of ocean circulation.

Climate modellers and physicists will no doubt scrutinize this new model carefully because it proposes a new way of including surface temperature flux. Most previous ocean modelling studies have used a 'temperature-restoring' boundary condition that forces ocean temperatures towards the temperature of the bottom of the atmosphere (in principle, after a correction for the difference between the downward flux of solar radiation and the upward flux of longwave radiation and latent heat; in practice, often ignoring this

of this basin by nutrient-depleted waters from near the surface of the ocean. Palaeo-oceanographic studies from the early 1980s onwards (see for example ref. 7) established that during the Last Glacial Maximum (LGM), nutrient concentrations in the deeper waters of the North Atlantic were higher than today, whereas in the upper deep and intermediate waters they were scarcer. Following the logic of Wust¹³, one would infer that a water mass spread from the North Atlantic in the LGM at shallower depths than found in the modern North Atlantic.

In fact, a glacial Atlantic circulation dominated by a cold intermediate/upper deep water formed in the northern North Atlantic and/or Labrador sea has been suggested previously^{14–16}. This situation obviously does not describe a 'conveyor off' mode. And in a forthcoming issue of *Nature*, Fichefet *et al.*¹⁷ show that palaeo-oceanographically inferred glacial surface boundary conditions driving their zonally averaged ocean model lead similarly to a weakened shallower conveyor.

Although these similarities between model results and palaeo-oceanographic data are encouraging, there is a cautionary lesson here. A simple change from one physically unrealistic parameterization to another perhaps slightly less unrealistic caused sweeping changes in the behaviour of the model. As we advance our understanding through a series of progressively more realistic models, how many more new model behaviours will we discover, and are any of them truly related to the behaviour of the real ocean? □

Ed Boyle is in the Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. Andrew Weaver is in the School of Earth and Ocean Sciences, University of Victoria, P.O. Box 1700, Victoria, British Columbia, Canada V8W 2Y2.

1. Rahmstorf, S. *Nature* **372**, 82–85 (1994).
2. Chamberlin, T. C. *J. Geol.* **14**, 363–373 (1906).
3. Stommel, H. *Tellus* **13**, 224–230 (1961).
4. Marotzke, J. in *Ocean Processes in Climate Dynamics: Global and Mediterranean Examples* (eds Malanotte-Rizzoli, P. & Robinson, A. R.) 79–109 (Kluwer, Amsterdam, 1994).
5. Broecker, W. S., Peteet, D. M. & Rind, D. *Nature* **315**, 21–26 (1985).
6. Bryan, F. O. *Nature* **323**, 301–304 (1986).
7. Manabe, S. & Stouffer, R. J. *J. Clim.* **1**, 841–866 (1988).
8. Bretherton, F. P. *Prog. Oceanogr.* **11**, 93–129 (1982).
9. Send, U. & Marshall, J. J. *phys. Oceanogr.* (in the press).
10. CLIMAP, *Seasonal Reconstructions of the Earth's Surface at the Last Glacial Maximum* (Geol. Soc. Am. Map and Chart Series, 1981).
11. Duplessy, J.-C. *et al. Oceanogr. Acta* **14**, 311–324 (1991).
12. Duplessy, J.-C. *et al. Paleoceanography* **3**, 343–360 (1988).
13. Wust, G. *The Stratosphere of the Atlantic Ocean*, 1–112 (Amerind, New Delhi, 1935).
14. McIntyre, A. *et al. in Investigations of Late Quaternary Paleoceanography and Paleoclimatology Memoir* (eds Cline, R. M. & Hays, J. D.) 43–76 (Geol. Soc. Am., 1976).
15. Boyle, E. & Keigwin, L. D. *Nature* **330**, 35–40 (1987).
16. Imbrie, J. *et al. Paleoceanography* **7**, 701–738 (1992).
17. Fichefet, T. *et al. Nature* (in the press).

The economics of extinction

Robert M. May

FROM a narrowly economic viewpoint, the best long-term harvesting strategy for biological populations with relatively low growth rates is to liquidate the stock. That conclusion, which assumes discounts on future returns, is based mainly on analyses involving a deterministic world. But real populations are of course subject to all manner of fluctuations. On page 88 of this issue¹ Lande *et al.* extend Colin Clark's pioneering work² to include stochastic population dynamics in the calculations. Under many circumstances, they find that

whales, tropical hardwoods and other slow-growing resources. Elephants and their ivory arguably fall into this category; ideally the optimal sustainable strategy may be to harvest ivory from elephants only after they die from natural causes^{4,5}. For most fish, or fast-growing softwoods, the problems are 'sole owner' ones (not that these are simple)².

Real populations exhibit fluctuations caused by external environmental uncertainties and by demographic events ('environmental and demographic stochasticity', in the jargon of the trade⁶). These undercut any simple conclusions based on deterministic population models. The effects of such stochasticity have been studied in harvesting models which ignore the long-term complications associated with discounting the future^{7,8}, that is, in the limiting case $\delta = 0$, which makes sense for those resources, such as many fish populations, where $r \gg \delta$. Population fluctuations are recognized as reducing the maximum average harvest below the deterministic 'maximum sustainable yield' (MSY). Furthermore, nonlinear population dynamics may interact with harvesting to cause larger fluctuations as harvesting pressures increase. These insights, reinforced by the collapse of many fisheries, have resulted in older MSY policies being replaced by more cautious management strategies^{9,10}.

Lande *et al.* emphasize that, once environmental and demographic stochasticities are introduced, eventual extinction will be the fate of any population. Different harvesting regimes will affect the average time to extinction, T , in different

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Grim harvest — ivory stacked for burning, Nairobi.

the result of incorporating such stochasticity into harvesting strategies that maximize the discounted 'present value' of the resource is faster extinction of the target population.

In deterministic harvesting models, the important thing is the population's intrinsic growth rate, r , in relation to the economists' inflation-corrected discount rate, δ (often taken to be around 5–7 per cent). If $r > \delta$, then management for sustainable harvests does make economic sense, provided there is an effective 'sole owner'. In the absence of an effective regulatory body, the familiar phenomenon of 'dissipation of economic rent' (or, more poetically, the 'tragedy of the commons'³) leads to overexploitation.

Less well recognized by most economists is the opposite case, when $r < \delta$. Here the economically optimal harvesting strategy is to cash in the entire stock (or at least harvest to the point where the last small remnant is too expensive to collect), and invest the realized capital where returns are higher. Thus purely economic considerations suggest liquidating large

the result of incorporating such stochasticity into harvesting strategies that maximize the discounted 'present value' of the resource is faster extinction of the target population.

Different harvesting regimes will affect the average time to extinction, T , in different

ent ways. Lande *et al.* prove an elegant theorem: for $\delta = 0$, the strategy that maximizes T has a harvesting rate zero when the population is below its unharvested stable equilibrium point (denoted by the 'carrying capacity', K), and as high as possible when the population has fluctuated above K . This is in remarkable contrast to the typical deterministic MSY strategy, which is to harvest at such a rate as to keep the population around some fraction of K ($K/2$ in the simple logistic metaphor). The new management procedures that the International Whaling Commission has been debating for the past several years have some of the character of Lande and colleagues' 'on-or-off' rule, but tend to have a set-point not at K , but somewhat below it. These management procedures, moreover, aim to recognize the uncertainties in estimating such quantities as K (ref. 10).

When stochastic dynamics are combined with non-zero economic discount rates, Lande and co-workers' conclusions are disturbing, in two ways.

First, even for discount rates below the critical rate, $\delta^* = r$, which spells doom for the stock, the population is in trouble. Maximizing the discounted present value of the harvest, even for $\delta < \delta^*$, can greatly reduce average times to extinction, as well as greatly reducing the real (undiscounted) harvest. The additional practical problem is that even modest values of δ can reduce the present value of the harvest to levels such that it is cheaper to liquidate the resource than to bear the cost of maintaining harvesting operations for the small added long-term profits. Hence estimates of critical discount rates, based on deterministic population dynamics, may often be optimistically high.

Second, the critical discount rate, δ^* , may itself be lower in stochastic models. For the logistic case, δ^* is the same for deterministic and stochastic models. But if the population exhibits 'depensation', such that net population growth is negative below some lower unstable point, then the critical discount rate, below which economic rationality suggests liquidation, will be lower than the

deterministic estimate.

What is called 'overexploitation' is the cited cause of endangerment for about one-third of the plants and animals on Red Lists¹¹. The work of Lande *et al.* is important in showing that levels of exploitation higher than previously expected, including deliberate harvesting to extinction, may be dictated by economic analyses which allow for stochastic effects. Optimal policies depend on how we assess the discount rate; δ is a political and social parameter, not just an economic one¹². An accountant may take δ to be the inflation-corrected bank discount rate. But a conservation biologist, or a politician looking to provide for the future (if

such there are), may be more inclined, in effect, to put $\delta = 0$.

This is what Lande *et al.* urge. But the practical problems are great. Even in the comparatively happy situation where $r > \delta$, problems of implementing effective sole ownership through regulatory bodies are considerable, particularly in traditionally overcapitalized and overstaffed industries², as the recent Fish Wars show. When $r < \delta$, and economic motives impel towards extinguishing the resource, things are correspondingly more difficult. □

Robert M. May is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

GEOPHYSICS

Overturning mantle models

William M. White

MOTIONS of lithospheric plates are certainly due to convection in the Earth's mantle. But how does the mantle convect? Do convection cells extend through its whole 2,900-km depth or is convection broken into two or more layers? These alternatives have incited many fierce and acrimonious debates among mantle specialists over the years. However, recent geodynamic modelling suggests a compromise: convection could alternate between whole-mantle and layered modes (see for example ref. 1). On page 63 of this issue, Stein and Hofmann² argue that alternating mantle convection modes may explain some enigmatic aspects of mantle and crustal evolution.

The Earth's crust has undoubtedly formed through time by additions from the mantle, but just how this growth has occurred is still problematic. At present, new crust is mainly added by volcanism associated with subduction zones: the Andes are an excellent example. Geologists have been inclined to apply the uniformitarian principle ('the present is the key to the past') and assume that this has been true throughout geological history.

Of late, though, studies have found that the rocks of some ancient terranes do not bear the chemical fingerprints of subduction zone volcanism. The Abitibi belt of Canada is a good example. This huge block of crust formed 2,700 million years ago in a mere 20 million years. Many, and perhaps most, of the rocks in the Abitibi belt lack the subduction zone's geochemical signature³. Such studies have reopened the question of how new crust forms. Furthermore, the current rate of Andean-type crustal growth is too slow to have produced the continents in the time available.

Many geologists now suspect that large igneous provinces like the Abitibi were

produced by the initial stages of mantle plumes. When mantle plumes first form, they are thought to have a large bulbous head. As the head approaches the surface, it partially melts, producing an episode of volcanism with extraordinary eruption rates. Volcanism associated with plume 'tails', such as Hawaii, is almost trivial by comparison. There are no current examples of this 'flood basalt' volcanism (which might be just as well), but several of these events occurred in the Cretaceous period (about 120 million years ago), producing the Ontong-Java Plateau in the western Pacific, the Kerguelen Plateau in the Indian Ocean and other such features.

Recognition that the Cretaceous was a period of extensive volcanism⁴ has given rise to speculation that there might be episodes when many plumes are initiated, and this is central to the thesis of Stein and Hofmann. They argue that the mantle alternates between periods of two-layer convection, when there are no plumes initiating, and periods of whole mantle convection when many plumes start to grow, leading to a burst of crustal growth. They point out that the geochronological record suggests that periods of rapid crustal growth alternate with longer periods, such as the present, in which crustal growth is slow.

One line of evidence that the crust formed from the mantle is that the upper mantle is depleted in just those elements, called incompatible elements, that are most concentrated in the crust. Radiogenic isotope variations (in other words, variations in isotopic composition that result from radioactive decay) provide some insights into this depletion process. Neodymium-143, produced by α -decay of samarium, has proved to be useful in this respect. Neodymium isotope ratios of the mantle appear most consistent with the

1. Lande, R., Eugen, S. & Saether, B. E. *Nature* **372**, 88–90 (1994).
2. Clark, C. W. *Mathematical Bioeconomics* 2nd edn (Wiley, New York, 1990).
3. Hardin, G. *Science* **131**, 1292–1295 (1960).
4. Milner-Gulland, E. J. & Beddington, J. R. *Proc. R. Soc. Lond.* **B252**, 29–37 (1993).
5. Basson, M., Beddington, J. R. & May, R. M. *Math. Biosci.* **104**, 73–95 (1991).
6. May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, 1973).
7. Beddington, J. R. & May, R. M. *Science* **197**, 463–465 (1977).
8. Reed, W. J. *Math. Biosci.* **41**, 273–307 (1978).
9. Ludwig, D., Hilborn, R. & Walters, C. *Science* **260**, 17–36 (1993).
10. Rosenberg, A. A., Fogarty, M. J., Sissenwine, M. P., Beddington, J. R. & Shepherd, J. G. *Science* **262**, 828–829 (1993).
11. Groombridge, B. (ed.) *Global Biodiversity* (Chapman & Hall, London, 1992).
12. May, R. M. *Nature* **263**, 91–92 (1976).

entire crust having formed within the first 500 million years or so of Earth history, but this is not consistent with the geochronological evidence for continual continent growth. This inconsistency was taken to indicate that upper mantle must be resupplied with incompatible elements by recycling of continental crust into the mantle (through subduction)⁵.

Patchett and Chauvel⁶ noted that the neodymium isotope data could be equally well explained if the upper mantle is resupplied with incompatible elements from below. Subsequent work by Galer and O'Nions⁷ found that the present ratio of thorium to uranium in the upper mantle is lower than the time-integrated value recorded by lead isotope ratios. They argued that lead now in the upper mantle cannot have been there long. Although subducted continental material has an appropriate neodymium isotopic composition to resupply the upper mantle, it does not have the proper lead isotopic composition. Mantle plumes, on the other hand, have both an appropriate neodymium and an appropriate lead isotopic composition⁸.

In the model that Stein and Hofmann are proposing, during periods of two-layer convection the upper mantle undergoes progressive depletion in incompatible elements through formation of oceanic crust and subduction-zone volcanism. During periods of whole-mantle convection, plumes rising from the lower mantle, which is less depleted in incompatible elements, resupply the upper mantle with these elements.

So the mantle overturn model of Stein and Hofmann may provide an explanation for two longstanding problems: episodic crustal growth and upper-mantle isotope compositions. On the other hand, there is more speculation here than hard evidence. There is much work yet to be done and many questions still to be answered before crust and mantle evolution can be understood. Perhaps most important, has crustal growth been truly episodic or does the distribution of ages merely reflect a preservation or sampling bias? What is the dominant mode of crust formation? Although the consensus of specialists may be swinging away from subduction volcanism and towards plume volcanism, that consensus has shifted before and may do so again.

And does the mantle really switch convective modes, or is this merely a figment

of some computer's imagination? The answer to that lies in understanding the thermodynamics of the mantle transition zone at 400 to 650 km depth. Phase changes in this region have the potential to retard the sinking of lithosphere and rise of mantle plumes. If mantle overturn occurs, these phase changes are likely to

be the toggle. Francis Birch remarked long ago that the transition zone is the key to understanding evolution of the mantle. That still seems to be true. □

William M. White is in the Department of Geological Sciences, Cornell University, Ithaca, New York 14853, USA.

IMMUNOLOGY

Death in the thymus

Ken Shortman and Roland Scollay

MOST T lymphocytes are born in the thymus. On page 100 of this issue¹ Surh and Sprent show that most new-born T-lineage cells also die there. They thereby settle a long-standing controversy.

The young thymus regenerates about one-third of its lymphoid content each day, a rate far in excess of the level needed to replenish the peripheral T-cell pools²⁻⁴. Most of this excess production consists of small CD4⁺8⁺ thymocytes generated in the cortex of the thymus. The rate of production of mature CD4⁺8⁻ and CD4⁻8⁺ medullary thymocytes is much lower and is more closely in balance with the rate of export of mature T cells⁴. So what happens to the excess production of cortical thymocytes? Surh and Sprent have tackled the question by using a new method for detecting DNA-strand breaks in cells in tissue sections, and have come up with direct evidence that the excess cells die *in situ* by the process of apoptosis and are cleared by thymic macrophages.

In 1966 Metcalf and colleagues^{2,3} discovered that the thymus was not a storehouse for long-lived memory lymphocytes, as they had suspected, but rather a site of rapid lymphopoiesis. Their balance sheets on lymphocyte turnover led them to the startling conclusion that virtually all lymphocytes born in the thymus must die there. But Nossal⁵, Weissman⁶ and others provided evidence that some cells left the thymus, a finding more in accord with its immunological function⁷. The development of a technique for direct labelling of thymocytes with a fluorescent dye⁸ provided a tag on thymic emigrants and allowed quantification of the export rate: each day, approximately 1% of all thymocytes, or 3% of the daily production, left the organ. This made the Metcalf estimates 97% correct. Even if the few lymphocytes that emigrated were crucial to function, most of the large daily production of thymocytes still appeared to die before departure.

The problem was that, by conventional histology, the thymus did not look like a cell graveyard. Although the incidence of visible dead cells would depend on the efficiency of the phagocytic system in

disposing of the remains, many histologists were unconvinced.

It did, at one stage, seem that a biochemical approach had demonstrated massive cell death within the thymus^{9,10}. Thymidine is incorporated much more efficiently into DNA than its analogue, iododeoxyuridine. If thymocyte DNA is labelled with these precursors, local cell death and reuse of DNA-derived nucleosides should lead to a relative retention of ³H-labelled thymidine but a relative loss of ¹²⁵I-labelled iododeoxyuridine. This was indeed observed. But the results lost their impact when the main reason for the loss of ¹²⁵I was found to be the chemical process of deiodination, rather than selective reuse of the non-iodinated nucleotide¹¹. In a reconsideration of these issues, Rothenberg¹² suggested the excess cortical thymocytes must leave the thymus to die elsewhere, although no extrathymic cemetery was identified.

Surh and Sprent have now shifted the search back to the thymus itself. In their model, thymic macrophages must be capable of recognizing apoptotic thymocytes by subtle changes in their surface membrane, and then rapidly digesting them, well before a histologist would recognize a cell death. Their count of 0.5–1% cortical thymocytes undergoing apoptosis allows an estimate of the total number dying *in situ*. If an apoptotic cell is visible by their technique for say one hour, 12–24% of thymocytes would be undergoing apoptosis each day. Because there is general agreement that apoptosis is very rapid, the data are consistent with all the excess thymus production being disposed of within the thymus.

In contrast to the situation in 1966, immunologists today have a rationale for extensive cell death within the thymus. The repertoire of T-cell antigen-receptor (TCR) specificity is initially generated by a random process of gene rearrangement. Stringent cell selection is then required to ensure that only T cells of appropriate specificity leave the organ. Such selection is of two types, positive and negative.

Positive selection¹³ is the rescue from programmed cell death of thymocytes with appropriate affinity for self mol-

1. Machetel, P. & Wever, P. *Nature* **350**, 55–57 (1991).
2. Stein, M. & Hofmann, A. *Nature* **372**, 63–68 (1994).
3. Vervoort, J. D., White, W. M., Thorpe, R. I. & Franklin, J. M. *Econ. Geol.* **88**, 1598–1614 (1993).
4. Larsen, R. L. *Geology* **19**, 547–550 (1991).
5. De Paolo, D. J. *Geophys. Res. Lett.* **10**, 705–708 (1983).
6. Patchett, P. J. & Chauvel, C. *Geophys. Res. Lett.* **11**, 151–153 (1984).
7. Galer, S. J. & O'Nions, R. K. *Nature* **316**, 778–782 (1985).
8. White, W. M. *Earth planet. Sci. Lett.* **115**, 211–226 (1993).

ecules of the major histocompatibility complex (MHC). These cells are hence eventually able to interact with others of the immune system of that individual. Cells that are not positively selected die of 'neglect'. Negative selection¹⁴ is the active elimination of thymocytes with a potentially destructive self-reactivity. The relative contribution of cell death by neglect or by active elimination had been unclear. By studying mutant mice lacking MHC molecules, Surh and Sprent identify failure of positive selection as the main reason for thymocyte death in the thymic cortex. They show that negative selection makes a detectable contribution, although with the model self-antigen they employ, cell death then takes place in the medulla.

Do we now understand the basis for this balance between cell death in the thymus and cell departure from it? Metcalf (who dubs the immunologists' rationale for intrathymic death a 'religious explanation') has calculated that the production of granulocytes in bone marrow also far exceeds their rate of departure into the bloodstream (personal communication). Specificity selection is unlikely to apply to these cells. It is also noteworthy that in TCR-transgenic mice where negative selection is absent or minimal, and all thymocytes are potentially open to positive selection, departure of cells from the thymus is not markedly increased¹⁵. This suggests that factors other than immune specificity regulate cell output, an area needing investigation. But whatever the basis of thymocyte rejection, the concept that large numbers of unwanted cells are disposed of within the thymus by apoptosis and macrophage digestion now seems to be firmly established. □

Ken Shortman is at The Walter and Eliza Hall Institute of Medical Research, Melbourne, Victoria 3050, Australia. Roland Scollay is at the Centenary Institute of Cancer Medicine and Cell Biology, Locked Bag No. 6, Newtown, New South Wales 2042, Australia.

Risky apolipoprotein in brain

Michel Goedert, Warren J. Strittmatter and Allen D. Roses

INHERITANCE of the E4 allele of the gene encoding apolipoprotein E (*APOE*) is one of the main risk factors for late-onset Alzheimer's disease. Dozens of studies have confirmed the original finding, but the mechanisms underlying the link are not yet clear. On page 92 of this issue¹, Ma *et al.* report that apoE isoforms promote β -amyloid fibril formation *in vitro*, apoE-4 having the most pronounced effect. Together with a paper by Sanan *et al.*²,

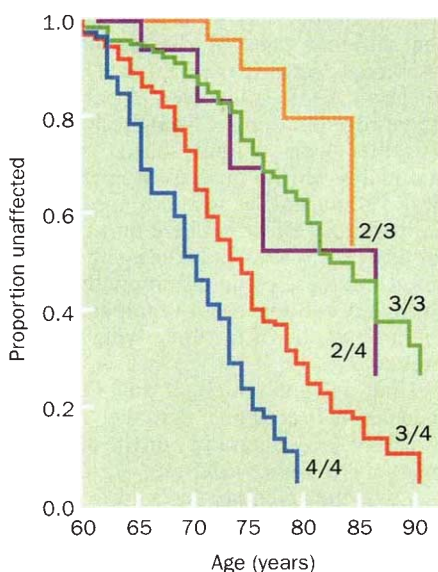
ing that APP is involved in the aetiology of at least some cases.

But the vast majority of Alzheimer's cases are of late-onset; most of these are believed to be sporadic, although some are familial. Linkage of some late-onset familial cases to chromosome 19q13.2, in conjunction with biochemical studies, has shown that inheritance of *APOE*-4, one of three common alleles, is a significant risk factor for these and sporadic instances of the disease⁹⁻¹². As a function of age, the risk increases with the number of *APOE*-4 alleles inherited, the mean age of onset being some 15 years earlier in individuals who inherit two E4 rather than two E3 alleles (see figure). Inheritance of *APOE*-2 decreases risk of the disease and increases age of onset¹³.

The apoE protein plays an important role in plasma cholesterol and triglyceride metabolism through its ability to bind lipids and to interact with cell-surface lipoprotein receptors¹⁴. The human protein (299 amino acids) is polymorphic — there are three major isoforms, apoE-2, apoE-3 and apoE-4, which differ by their cysteine and arginine contents at positions 112 and 158. The most common is apoE-3, which has a cysteine at position 112 and an arginine at position 158; apoE-4 has arginines at both positions; and apoE-2, the least common isoform, has cysteines. So a single amino-acid difference between apoE-3 and apoE-4 accounts for a big difference in the expression of the disease. The therapeutic implications are obvious.

Substantial amounts of apoE are found in the central nervous system, where it is chiefly made by astrocytes¹⁵. It is believed to be internalized by nerve cells following binding to the low-density lipoprotein receptor-related protein¹⁶. In the Alzheimer's brain, apoE is found in most amyloid plaques, vascular amyloid and in some nerve cells, with or without neurofibrillary lesions^{17,18}. Brain sections from patients homozygous for *APOE*-4 show more amyloid plaques and stronger A β immunoreactivity per plaque than sections from patients homozygous for *APOE*-3 (ref. 19). Moreover, *in vitro* A β binds more avidly to apoE-4 than to apoE-3 (ref. 20).

But what effects might the different apoE isoforms have on A β fibril formation? Ma *et al.*¹ have used chemically synthesized A β ₁₋₄₂, the principal component of amyloid deposits in Alzheimer's disease²¹, to show that, *in vitro*, apoE-4 has a marked stimulatory effect on both the rate of formation and the number of amyloid fibrils; the effect of apoE-3 is less marked, and that of apoE-2 smaller still.



Distribution of age of onset of Alzheimer's disease for subjects with each common *APOE* genotype. Onset curves were estimated by Kaplan-Meier product limit distributions. Data were limited to late-onset Alzheimer's disease and should not be generalized to all populations at risk — epidemiologically based distributions will be needed before risks can be estimated in the general population (figure reproduced from ref. 33).

which describes similar results, this work provides insights into interactions between apoE and amyloid- β protein (A β).

Alzheimer's neuropathology is characterized by abundant senile plaques and neurofibrillary lesions. Amyloid- β is the main component of extracellular amyloid fibrils^{3,4}, whereas the microtubule-associated protein tau in a hyperphosphorylated state is the main fibrous constituent of the intracellular neurofibrillary lesions^{5,6}. Familial forms with an early age of onset comprise 5–10% of all Alzheimer's cases, the major locus is at chromosome 14q24.3 (ref. 7), and several laboratories are working towards the identification of the culprit gene. A smaller proportion of early-onset familial cases is associated with point mutations in the gene encoding amyloid precursor protein (APP) on chromosome 21 (ref. 8), indicat-

1. Surh, C. D. & Sprent, J. *Nature* **372**, 100–103 (1994).
2. Matsuyama, M. B., Wiadrowski, M. N. & Metcalf, D. J. *exp. Med.* **123**, 559–576 (1966).
3. Metcalf, D. in *The Thymus: Experimental and Clinical Studies*, CIBA Fdn Symp. (eds Wolstenholme, G. W. E. & Porter, R.) 242–263 (Churchill, London, 1966).
4. Shortman, K., Egerton, M., Spangrude, G. J. & Scollay, R. *Sem. Immun.* **2**, 3–12 (1990).
5. Nossal, G. J. V. *Ann. N. Y. Acad. Sci.* **120**, 171 (1964).
6. Weissman, I. L. *J. exp. Med.* **126**, 291–304 (1967).
7. Miller, J. F. A. P. & Osoba, D. *Physiol. Rev.* **47**, 437 (1967).
8. Scollay, R., Butcher, E. & Weissman, I. *Eur. J. Immun.* **10**, 210–218 (1980).
9. Joel, D. D., Chanana, A. D., Cottier, H., Cronkite, E. P. & Laissue, J. A. *Cell Tissue Kinet.* **10**, 57–69 (1977).
10. McPhee, D., Pye, J. & Shortman, K. *Thymus* **1**, 151–161 (1979).
11. Quackenbush, R. C. & Shields, A. F. *Cell Tissue Kinet.* **21**, 381–388 (1988).
12. Rothenberg, E. V. *Immun. Today* **11**, 116–119 (1990).
13. von Boehmer, H. *Immun. Today* **7**, 333–336 (1986).
14. Kappler, J. W., Roehm, N. & Marrack, P. *Cell* **49**, 273–280 (1987).
15. Kelly, K. A., Pircher, H., von Boehmer, H., Davis, M. M. & Scollay, R. *Eur. J. Immun.* **23**, 1922–1928 (1993).

These results may explain the large number of A β deposits in Alzheimer's patients homozygous for APOE-4. Soluble A β is constitutively produced as part of normal APP metabolism, with A β ₁₋₄₀ as a majority species and A β ₁₋₄₂ as a minority species^{22,23}. The mechanisms that lead from soluble A β in normal brain to amyloid fibrils in Alzheimer's disease brain are only poorly understood. An increased production of A β ₁₋₄₂ is the net effect of several APP mutations linked to early-onset forms of Alzheimer's^{24,25}, suggesting that it may be one mechanism that leads to A β deposition.

Ma *et al.* show that proteins associated with amyloid deposits in the Alzheimer's brain promote A β fibril formation *in vitro*. They report a similar effect to apoE for α_1 -antichymotrypsin, another protein associated with amyloid plaques. The relevance of apoE-induced stimulation of A β -fibril formation for understanding Alzheimer's disease depends on the part played by A β deposits in the characteristic nerve-cell degeneration that accounts for the disease symptoms. Amyloid- β protein can be toxic to cultured fetal or neonatal mammalian nerve cells²⁶, but the evidence for such an effect on adult nerve cells *in vivo* is scarce. Moreover, early neuropathological stages of Alzheimer's disease frequently show neurofibrillary lesions in the absence of A β deposits²⁷, suggesting that the deposits may not themselves trigger nerve-cell degeneration.

Pathological changes that are qualitatively indistinguishable from those in Alzheimer's disease are an almost invariable accompaniment of aging, but to a much lesser extent. From survival curves, it seems that, given a lifespan of 130–140 years, everybody would eventually develop the disease, whatever the APOE allele (see figure). The apoE-2 and apoE-3 isoforms may delay this process more than apoE-4 by prolonging the survival of affected nerve cells.

Although the experiments described here were carried out with A β , it should be noted that neurofibrillary lesions made of hyperphosphorylated tau protein correlate better with the disease symptoms than do amyloid plaques²⁸. Here it may be significant that apoE-3 binds to the microtubule-binding domain of tau *in vitro*, whereas apoE-4 shows no significant binding²⁹. Neither isoform binds to hyperphosphorylated tau. Light- and electronmicroscopic studies have shown apoE-like immunoreactivity in the cytoplasm of some human cortical neurons³⁰. These observations have led to the proposal that interactions between apoE-3 and tau may protect tau from hyperphosphorylation, and account for the different ages of disease onset in individuals with APOE3/3 and APOE4/4 alleles²⁹. ApoE is involved in peripheral nerve regeneration following lesioning³¹. Together with a lipid source, apoE-3 stimulates neurite outgrowth in cultured dorsal root ganglia, whereas apoE-4 has an inhibitory effect³². There may be similar differences at work in the aging human brain and these could differentially affect the ability of nerve cells to compensate for the progression of the underlying disease process.

Time will tell which, if any, of these mechanisms accounts for the striking effects of APOE genotype on the development of Alzheimer's disease. A great deal hangs on the outcome — a delay in the age of onset by 10–15 years, which might be induced by a drug acting like apoE-2, would result in a massive reduction in the number of disease cases. □

Michel Goedert is at the MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK. Warren J. Strittmatter and Allen D. Roses are in the Departments of Medicine (Neurology) and Neurobiology, Duke University Medical Centre, Box 2900, Durham, North Carolina 27710, USA.

Real recycling

An Indian city recently acquired an expensive Western waste-utilization plant, a furnace intended to power generators. It never worked. The diligent rubbish-sorters of the city scoured the rubbish as it came in, removing everything of even trivial value. What they left behind, mainly rotting vegetable matter, would hardly burn. The costly, token, feel-good recycling of the West was neatly subverted by the genuine recycling of the Indian rubbish-sorters.

And yet the West does support a little real recycling. The car-boot sale, the bring-and-buy stall and the traditional English jumble sale all return unwanted material to the economy. Significantly, they operate on individual objects, not by clumsy bulk reprocessing. Daedalus wants to modernize them.

He points out how modern communications have transformed the financial markets, and are now transforming the industrial ones. So DREADCO programmers are writing the software for a continuous national jumble sale, to be maintained on the Internet. Each user will advertise any or all of his surplus possessions on the Net, while simultaneously appealing for anything he happens to want.

Daedalus expects a massive trade to develop. At first it will be dominated by fairly large and expensive consumer durables, laboratory and industrial equipment and so on. Those who throw such things out for trivial faults, or routinely replace them with the latest model, will be efficiently matched to the handymen who can repair them or work with older apparatus. But soon smaller objects will come crowding in. Books of all kinds will be sought and exchanged; then raw materials. The boilermaker's useless offcut is valuable raw material to the toolroom engineer, and his scrap would be snapped up by the model maker. A surplus load of an industrial chemical would be cornucopia to the laboratory chemist. Such materials will repeatedly re-enter the economy in ever increasing subdivision, until they reach the point where the cost of moving them further outweighs the value of the trade.

Each user of the service will release his own 'gopher' programs into the Net to seek out and strike bargains with other users and their gophers. The automated haggling of thousands of individual trades will establish 'going rates' for the most surprising objects: half-used tins of paint in all shades, nuts and bolts of every size, lengths of cable, pipe and string and bin-ends of cement. This truly efficient market will be far more economical than the token recycling of a few raw materials.

David Jones

- Ma, J., Yee, A., Brewer, H. B. Jr., Das, S. & Potter, H. *Nature* **372**, 92–94 (1994).
- Sanan, D. A. *et al.* *J. clin. Invest.* **94**, 860–869 (1994).
- Glenner, G. G. & Wong, C. W. *Biochem. biophys. Res. Commun.* **120**, 885–890 (1984).
- Masters, C. L. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **82**, 4245–4249 (1985).
- Goedert, M., Wischik, C. M., Crowther, R. A., Walker, J. E. & Klug, A. *Proc. natn. Acad. Sci. U.S.A.* **85**, 4051–4055 (1988).
- Lee, V. M.-Y., Balin, B. J., Otvos, L. & Trojanowski, J. Q. *Science* **251**, 675–678 (1991).
- Schellenberg, G. D. *et al.* *Science* **258**, 668–671 (1992).
- Goate, A. *et al.* *Nature* **349**, 704–706 (1991).
- Pericak-Vance, M. A. *et al.* *Am. J. hum. Genet.* **48**, 1034–1050 (1991).
- Strittmatter, W. J. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **90**, 1977–1981 (1993).
- Saunders, A. M. *et al.* *Neurology* **43**, 1467–1472 (1993).
- Corder, E. H. *et al.* *Science* **261**, 921–923 (1993).
- Corder, E. H. *et al.* *Nature Genet.* **7**, 180–184 (1994).
- Weisgraber, K. H. *Adv. Prot. Chem.* **45**, 249–302 (1993).
- Boyles, J. K., Pitas, R. E., Wilson, E., Mahley, R. W. & Taylor, J. M. *J. clin. Invest.* **76**, 1501–1513 (1985).
- Rebeck, G. W., Reiter, J. S., Strickland, D. K. & Hyman, B. T. *Neuron* **11**, 575–580 (1993).
- Namba, Y., Tomonaga, M., Kawasaki, H., Otomo, E. & Ikeda, K. *Brain Res.* **541**, 163–166 (1991).
- Han, S. H. *et al.* *Exp. Neurol.* **128**, 13–26 (1994).
- Schmechel, D. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **90**, 9649–9653 (1993).
- Strittmatter, W. J. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **90**, 8098–8102 (1993).
- Iwatsubo, T. *et al.* *Neuron* **13**, 45–55 (1994).
- Haass, C. *et al.* *Nature* **359**, 322–326 (1992).
- Shoji, M. *et al.* *Science* **258**, 126–129 (1992).
- Citron, M. *et al.* *Nature* **360**, 672–674 (1992).
- Suzuki, N. *et al.* *Science* **264**, 1336–1340 (1994).
- Kowall, N. W., Beal, M. F., Busciglio, J., Duffy, L. K. & Yankner, B. A. *Proc. natn. Acad. Sci. U.S.A.* **88**, 7247–7251 (1991).
- Braak, H. & Braak, E. *Acta Neuropathol.* **82**, 239–259 (1991).
- Arriaga, P. V., Growdon, J. H., Hedley-White, E. T. & Hyman, B. T. *Neurology* **42**, 631–639 (1992).
- Strittmatter, W. J. *et al.* *Proc. natn. Acad. Sci. U.S.A.* (in the press).
- Han, S. H. *et al.* *J. Neuropathol. exp. Neurol.* **53**, 535–544 (1994).
- Ignatius, M. J. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **83**, 1125–1129 (1986).
- Nathan, B. P. *et al.* *Science* **264**, 850–852 (1994).
- Roses, A. D. *et al.* *Lancet* **343**, 1564–1565 (1994).

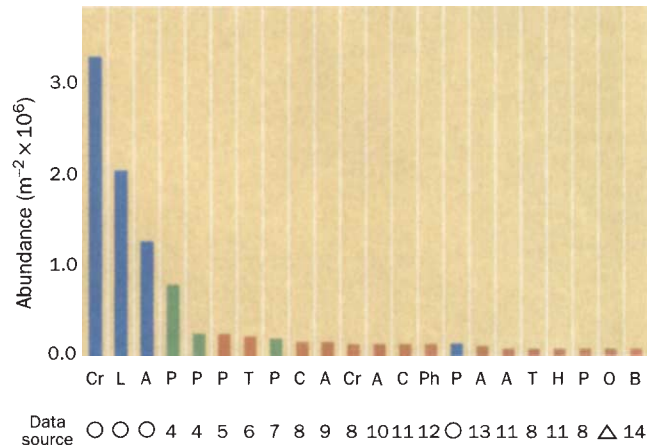
Hotspots of benthic production

SIR — Topographical features that accumulate organic debris, such as submarine canyons, are common along many coasts and can support hotspots of secondary production. The floor of the La Jolla submarine canyon (32° 52' N, 117° 15' W) is covered by a persistent mat of surfgrass and kelp detritus from a depth of 15 m to at least 300 m. The detritus is inhabited by a dense assemblage of amphipod (*Orchomene limodes*, *Aoroides spinosus*) and leptostracan (*Nebalia* spp.) crustaceans which at times achieve densities of more than 3 million individuals and biomass exceeding 1 kg (dry weight) per square metre. In monthly samples collected from March 1992 to March 1993, leptostracan density averaged $6.9 \times 10^5 \text{ m}^{-2}$ (s. d. 4.4×10^5) and amphipods $7.8 \times 10^5 \text{ m}^{-2}$ (s. d. 2.7×10^5). The combined maximum density of these animals was $3,240,000 \text{ m}^{-2}$, an order of magnitude greater than any natural macrofaunal assemblage reported in the literature (see figure).

The leptostracans averaged 4.5 mm long, and their secondary production, as calculated by the cohort summation of losses method¹, was $3,300 \text{ g dry weight m}^{-2} \text{ yr}^{-1}$ (production/biomass ratio = 7.2). This value is an underestimate because many of the largest size classes could not be assigned to a cohort and were omitted from the calculation. The mean dry biomass of the amphipod assemblage during the study was 168 g m^{-2} and, assuming a production/biomass ratio of 4.0, amphipod annual production is estimated as $672 \text{ g dry weight m}^{-2} \text{ yr}^{-1}$ (average length of amphipods, 2.2 mm). I found only two measurements of secondary production in the literature with values approaching those found in the

present study (both were for bivalves).

Although enormous, the secondary production in the La Jolla canyon is localized, in contrast with the ampeliscid amphipods of the northern Bering Sea, which have much lower productivity per unit area ($30\text{--}40 \text{ g dry wt m}^{-2} \text{ yr}^{-1}$); but that estimate applies to an area of 2.2×10^4 to $3.75 \times 10^4 \text{ km}^2$ (ref. 2), a scale that may be unsurpassed anywhere in the world. During this study more than



Density (m^{-2}) of the detritus-mat invertebrates compared with reports of dense assemblages in the literature. First three columns, densities of total crustaceans (Cr), leptostracans (L) and amphipods (A) from this study. Blue, data from this study. Red, data from polluted environments. Green, data from other, undisturbed, environments. Additional abbreviations: P, polychaete; C, community; Ph, phoronid; T, tanaid; H, harpacticoid copepod; O, oligochaete; B, bivalve. Open circles, data from this study; triangle, J. Crooks, unpublished data; numbers, citations for the above data.

$2,000 \text{ m}^2$ of the detritus habitat were quantitatively sampled, but the total area of the mat within the La Jolla submarine canyon system probably exceeds $100,000 \text{ m}^2$. The dense assemblage of crustaceans and their fish predators have been observed as deep as 60 m, the maximum depth examined.

The amphipods are generally distributed within the top 10–20 cm of the detritus-mat and the leptostracans are most abundant 5–25 cm into the detritus; however, gut-content analysis and laboratory investigations indicate that leptostracans are the preferred prey of the mat-associated fishes. Large numbers of both demersal and pelagic fishes are nearly always present over the mat, and those fish, even species that typically prey on plankton, mega-invertebrates, and other fish, generally have their guts filled with the detritus-dwelling crustaceans. Fish are the tertiary predators in this system and functionally 'shunt' the detritus-based saprotrophic production into the classic biotrophic coastal food web (G. Polis and D. Strong, manuscript submitted). Most of the fishes common at 15 m are also abundant at 60 m; several of these species were found

below their maximum reported depth, and probably use the habitat and its resident fauna at much greater depths in the canyon.

These results raise the question of the general importance of such localized food hotspots to fish populations. At first glance they would seem a modest contribution in relation to the overall shelf and slope habitat, but consideration of the oceanographic literature suggests that hotspots may be critical for many species. The extreme importance of high-density food patches to many species of pelagic animals is becoming increasingly clear. The fish and zooplankton that survive and reproduce have somehow been able to locate and exploit food-rich patches. The same is probably true for many species of benthic and demersal fish and invertebrates. Cold seep and hydrothermal vent communities occur on a much smaller scale and are temporally ephemeral, but they also support high biomass and presumably very high rates of production, enriching what would otherwise be a very food-poor environment.

Productivity hotspots may, in aggregate, be an important energy supply for fish production along some coasts. Systems such as this provide an important mechanism to channel marine macrophyte production into higher trophic levels through a saprotrophic food web. If, as suggested by Pimm³, extensive omnivory is destabilizing, extremely high secondary production by omnivores may persist only in detritus-based food webs with large inputs of allochthonous production. This situation exists in the La Jolla canyon and is expected to occur in other submarine canyons near to populations of marine macrophytes.

Eric W. Vetter

Scripps Institution of Oceanography,
University of California, San Diego,
La Jolla, California 92093-0201,
USA

Scientific Correspondence

Scientific Correspondence is a relatively informal section of *Nature* in which matters of general scientific interest, not necessarily those arising from papers appearing in *Nature*, are published. Because there is space to print only a small proportion of the letters received, priority is usually given according to general interest and topicality, to contributions of fewer than 500 words, and to contributions using simple language.

1. Crisp, D. J. in *Methods for the Study of the Marine Benthos* (eds Holme, N. A. & MacIntyre, A. D.) 284–372 (Blackwell, Oxford, 1984).
2. Highsmith, R. C. & Coyle, K. O. *Nature* **344**, 862–864 (1990).
3. Pimm, S. L. *Food Webs* (Chapman and Hall, London, 1982).
4. Grassle, J. F. & Grassle, J. P. *J. mar. Res.* **32**, 253–284 (1974).
5. Levin, L. *Ecology* **65**, 1185–1200 (1984).
6. Highsmith, R. C. *Ecology* **63**, 329–337 (1982).
7. Anger, K. *Int. Rev. ges. Hydrobiol.* **62**, 245–254 (1977).
8. Dayton, P. K. & Oliver, J. *Science* **197**, 55–58 (1977).
9. Birklund, J. *Ophelia* **16**, 187–203 (1977).
10. Oliver, J. S., Slattery, P. N., Silberstein, M. A. & O'Connor, E. F. *Can. J. Zool.* **62**, 41–49 (1984).
11. Baden, S. P. *Neth. J. Sea Res.* **27**, 81–92 (1990).
12. MacGinitie, G. E. & MacGinitie, N. *Natural History of Marine Animals* (McGraw-Hill, New York, 1968).
13. Avery, W. E. & Hawkinson, C. *Northwest Sci.* **66**, 199–203 (1992).
14. Gardner, J. P. A. & Thomas, M. L. H. *Mar. Ecol. Prog. Ser.* **39**, 31–36 (1987).

Spatiotemporal chaos

SIR — Egolf and Greenside¹ recently investigated the relation between dimension density and correlation length in an extended spatiotemporally chaotic system. For spatiotemporally chaotic systems the fractal dimension, D , of the attractor typically grows proportionally to the volume of the system, making the dimension density dD/dV a more appropriate quantity. The reason is that the chaotic fluctuations have a finite correlation length ξ , so distant parts behave almost independently. Assuming that ξ is the only important length scale, it is natural to expect the dimension density times the correlation volume (the total number of degrees of freedom inside each correlation volume) to be constant. In other words (for a one-dimensional system)

$$dD/dL \approx 1/\xi \quad (1)$$

Egolf and Greenside¹ computed dimension densities and correlations for the complex Ginzburg–Landau equation

$$\partial_t u(x, t) = u + (1 + ic_1) \partial_x^2 u - (1 - ic_3) |u|^2 u \quad (2)$$

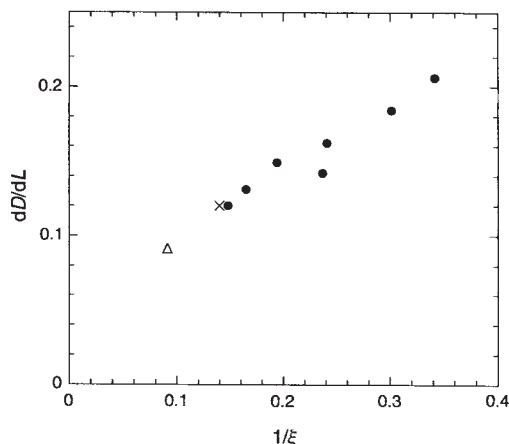
which describes the dynamics of a complex field $u = re^{i\phi}$ in a one-dimensional medium. (This equation can, for example, describe an oscillatory chemical reaction; the complex amplitude u is the Fourier component corresponding to the temporal period.)

The simulations were carried out in a regime where it is believed² that the nature of the chaotic dynamics changes from ‘defect’ to ‘phase’ turbulence. Approaching the transition from the defect turbulence side, the correlation length ξ was found to grow rapidly, and thus one should expect equation (1) to hold, but this turned out not to be so. The rapid variations in ξ are not reflected in dD/dL . This discrepancy was left by Egolf and Greenside as a provoking open question.

To understand what goes wrong, one has to understand the nature of the transition occurring in equation (2). In the defect-dominated state, the phase of the complex field u is not globally well defined, since it jumps by $\pm 2\pi$ around each defect. In the phase-turbulent regime, however, the defects have little or no role. The phase field becomes continuous and is expected to be well approximated by the Kuramoto–Sivashinsky (KS) equation³:

$$\partial_t \phi(x, t) \approx -v \partial_x^2 \phi - \mu \partial_x^4 \phi - \lambda (\partial_x \phi)^2, \text{ with } v = (c_1 c_3 - 1), \mu = \frac{1}{2} c_1^{-2} (1 + c_3^2) \text{ and } \lambda = (c_1 + c_3)$$

(although note that the nonlinear terms which are left out can effectively renormalize these values). The linear instability of the KS equation creates a cell structure with typical cell sizes $l_c = 2\pi \sqrt{2\alpha}$ with $\alpha = \sqrt{\mu/\nu}$ and from the work of Manneville⁴ the dimension density is known to be close to 2 per cell, or $dD/dL \approx 0.23/\alpha$, which, for the parameter values close to



Dimension density dD/dL against inverse information length ξ^{-1} for the complex Ginzburg–Landau equation. Our simulations had $L = 1,000$ and used a pseudospectral method. Information functions were computed from the evolution of $u(x, t)$ over 1,310,720 time units. The values of dD/dL are from ref. 1; the points with $dD/dL < 0.15$ are in the phase chaos regime; those with $dD/dL > 0.15$ correspond to defect chaos. See text for discussion of symbols.

the transition studied by Egolf and Greenside, is of the same order as the dimension density for the complex Ginzburg–Landau equation. Thus the cellular phase structure should introduce a new relevant length scale of the order of l_c . Further, equation (1) is trivially satisfied if we take l_c as the correlation length ξ . This can be seen by rescaling $x = \alpha x', t = \beta t'$ and $\phi = \gamma \phi'$ with $\alpha = \sqrt{\mu/\nu}$, $\beta = \mu/\nu^2$, and $\gamma = \nu/\lambda$ which makes all the coefficients ν , μ and λ unity.

The correlation function studied in ref. 1 is $C(x-x') = \langle u(x, t) u^*(x', t) \rangle$. This function is appropriate for the defect-dominated phase, but rather insensitive to the phase fluctuations on the scale l_c . From correlations of $r(x, t)$, and of $\partial_x \phi(x, t)$ in the phase turbulence regime, we can pick out a shorter length scale of the order of l_c . To avoid the troubles connected with oscillations of standard correlation functions, whose exponential decay is seldom very convincing in spatiotemporal chaos, we use so-called information functions^{5,6}, which express the correlations between the fields $A(x, t)$ and $A(x', t)$ at two spatially separated points directly in terms of their joint probability distribution function. We find simple exponential decay of these functions, and this method allows us to compute coherence lengths ξ of the order of, or even below, l_c .

In the figure we show the dimension densities dD/dL from ref. 1 against our coherence length ξ . The dots correspond to information functions for the modulus r , varying c_3 (from left to right) between 0.7 and 0.9 with c_1 , fixed at 3.5. The cross and the triangle are found from information functions for $\partial_x \phi$ for the full complex Ginzburg–Landau equation and for the KS approximation, respectively, both at $c_3 = 0.7$. Note the closeness of all three quantities in the phase-turbulent regime ($c_3 = 0.7$), where they are of the order $l_c/2$. In the KS approximation, both ξ and $(dD/dL)^{-1}$ scale as α . Thus the plot for varying c_3 would be a straight line from the origin through the triangle, with equation (1) obviously satisfied in that approximation.

In contrast, the length scale computed from ref. 1, $C(x-x') = \langle u(x, t) u^*(x', t) \rangle$, should not satisfy equation (1). In the KS approximation, this corresponds to computing the correlation function of $u(x, t) = \text{const} \times e^{i\phi(x, t)}$, for which (following ref. 2) the resulting correlation length ξ_1 is expected to scale as α/γ^2 , being sensitive to scaling of the phase field. Thus, again in the KS approximation, which should become better as c_3 is lowered into the phase-turbulent regime, the product $dD/dL \cdot \xi_1 \rightarrow \infty$ as the

Benjamin–Feir transition ($c_3 = 1/c_1$) is approached.

Although we do not claim to have found a general recipe for computing correlations of the chaotic fluctuations, we believe that we have pointed to the relevant length scales in the present problem for which there is no contradiction with equation (1).

We have benefited from discussions with Henry Greenside during this work.

Tomas Bohr

*The Niels Bohr Institute,
Blegdamsvej 17,
2100 Copenhagen Ø,
Denmark*

Eric Bosch

Willem van de Water
*Physics Department,
Eindhoven University of Technology,
PO Box 513,
5600 MB, Eindhoven,
The Netherlands*

1. Egolf, D. A. & Greenside, H. S. *Nature* **369**, 129–131 (1994).
2. Shraiman, B. I. et al. *Physica D* **57**, 241–248 (1992).
3. Kuramoto, Y. *Chemical Oscillations, Waves and Turbulence* (Springer, Berlin, 1984).
4. Manneville, P. in *Propagation in Systems Far From Equilibrium* (eds Westfeldt, J. E. et al.) 265–279 (Springer, Berlin, 1988).
5. Fraser, A. M. & Swinney, H. L. *Phys. Rev. A* **33**, 1134–1140 (1986).
6. van de Water, W. & Bohr, T. *CHAOS* **3**, 747–756 (1993).

Standard weights in ancient Andes

SIR — New evidence suggests that a standardized measure of weight was used in the prehispanic Andes. Small, prehispanic beam balances have been found throughout the area^{1–5}, but were thought to have been used with internally consistent unique weight sets to measure proportions only. No standardization was thought to have existed^{6,7}. I have found that the weights of four copper ingots from various locations on the Peruvian coast correspond to regular multiples of a

ideal weights of 50 and 100 *H* (see table). This accuracy is about what one would expect for ingots made, as these were, by remelting a carefully weighed quantity of copper prills⁸. Some copper would have been lost in the formation of crucible slag. Some impurities, notably iron, accounting for a few per cent of the prills' weight, would have segregated and been removed with the slag. These variables would have made more precisely replicated ingot weights impossible. The two Cerro Huaranga ingots, found together, could reflect an ancient pairing to make up a mass of copper weighing exactly 150 *H*.

So far, only two additional copper

Together, these weights and ingots suggest the existence of a system of standardized weights used from Ecuador to Peru's south coast. A late prehistoric date is indicated by the late textile accompanying the south coast ingot, and the late occupation of Cerro Huaranga's smelting workshops (AD 1100–1535) and Chan Chan (AD 1000–1470). The Chan Chan ingot predates the Inca conquest of the Chimú Empire in AD 1470, and so the system was not an Inca innovation. While the extent and antiquity of the system are still obscure, its very existence offers new insights into control of commodities, organization of production and circulation of goods in the prehispanic Andes.

Stephen M. Epstein

Department of Anthropology,
University of Pennsylvania,
Philadelphia,
Pennsylvania 19104,
USA

Ingot	Metric weight (g)	Andean weight (<i>H</i>)	Approximation (<i>H</i> ± %)
<i>Cerro Huaranga</i>			
BG 80-1	182	47.9	50–4.2
BG 80-2	399	105	100 + 5.0
BG 80-1 + BG 80-2	581	153	150 + 1.9
<i>Chan Chan</i>			
AMNH B/5200	386	102	100 + 1.6
<i>La Lengua</i>			
Ref. 11	1,100	289	300 + 3.5

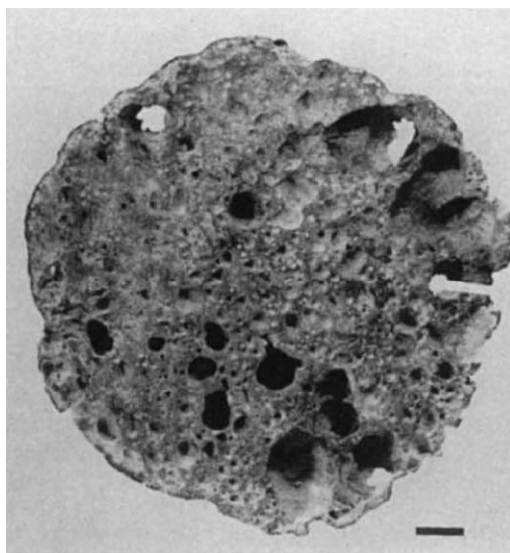
unit originally proposed by Nordenskiöld⁵ for prehispanic weights.

Most known Andean weights date to the Postconquest period and consist of simple pebbles carefully selected to correspond to the Spanish colonial system of *onzas* and *castellanos*. Nordenskiöld, however, identified 13 weights that did not conform to the Spanish system: one set of nine pebble weights in a cloth bag from Pachacamac, Peru; and four individual weights from various Andean locations. He found their weights to be multiples of a unit he called *H*, perhaps for *huarcu*, the word for weight in both Quechua and Aymará. The heaviest of the nine pebble set weighs 25 *H*; the other eight together weigh 25 *H*. The mean value of *H* is 3.80 ± 0.058 g.

Nordenskiöld's other weights, unfortunately, are vaguely provenanced and perhaps selectively chosen. Nevertheless, they include two pebbles ($H = 3.71$ and 3.89 g) from a "work-basket on the Peruvian coast", a somewhat oxidized lead weight ($H = 3.75$ g) from "Baranca", and a pierced quartz stone ($H = 3.88$ g) from Pinal, Ecuador. Including these additional finds would yield a mean *H* of 3.80 ± 0.067 g.

More recently, I found two plano-convex ingots of arsenical copper at the late prehistoric copper smelting site of Cerro Huaranga on Peru's north coast. These ingots approximate to

ingots that closely resemble the Cerro Huaranga ingots have been found. Both approximate to ideal weights, 100 and 300 *H*, respectively. The former, in the American Museum of Natural History^{9,10}, came from a grave in the Chimú capital of Chan Chan on Peru's north coast. Caley and Easby¹¹ acquired the latter on the south coast, at "La Lengua", in the Ingenio valley of Ica Province. It was wrapped in a warp-faced, plain-weave cotton cloth, with the alternating plain and barred stripes most common in textiles between AD 1200 and 1535.



A plano-convex ingot of arsenical copper found at Cerro Huaranga, Peru. The ingot is designated BG 80-2. Scale bar, 1 cm. (Photo by N. Hartmann.)

1. Nordenskiöld, E. *J. Soc. Am. Paris* **13**, 169–175 (1921).
2. Rostworowski de Diez Canseco, M. in *Actas del II Congreso Nacional de Historia* 1953 Vol. 2, 102–115 (Lima, Peru, 1960).
3. Baessler, A. *Ancient Peruvian Art* (transl. Keane, A.H.) Plate 161 (Dodd, Mead, New York, 1902–1903).
4. Schmidt, M. *Kunst und Kultur von Peru* 544 (Propyläen, Berlin, 1921).
5. Nordenskiöld, E. *Man* **30**, 215–221 (1930).
6. Bennett, W. C. in *Handbook of South American Indians* Vol. 5 (ed. Steward, J.) 601–610 (Smithsonian Institution, Washington, 1949).
7. Rostworowski de Diez Canseco, M. in *La tecnología en el mundo andino: Runakunap Kawsayninkupaq Rurasquankunaqa* (eds Lechtman, H. & Soldi, A. M.) 380–405 (Universidad Nacional Autónoma de México, México City, 1981).
8. Epstein, S. M. *MASCA J.* **2**, 57–62 (1982).
9. Meand, C. W. *Anthrop. Pap. Am. Mus. nat. Hist.* **12** (Pt 2) 16–52 (1915).
10. Petersen, G. G. *Minería y metalurgia en el antiguo Perú* (Museo Nacional de Antropología y Arqueología, Lima, 1970).
11. Caley, E. R. & Easby, D. T. Jr *Am. Antiq.* **25**, 59–65 (1959).

The 'St Jude' gambit

SIR — Oliver Goodenough and Richard Dawkins (*Nature* **371**, 23–24; 1994) have overlooked the fact that prior exposure to an appropriate mind virus may be a prerequisite for susceptibility to infection by another mind virus.

They themselves report the "experiencing of waves of mild, irrational anxiety on deciding not to comply" with the postal parasite's instructions. These symptoms disclose a prior infection, a 'meme', that was successfully implanted in them. It required a challenge from the St Jude virus to uncover the meme. Perhaps further taxonomic refinement will show that all memes are mind viruses, but not all mind viruses are memes.

Ian Dunn

48 Frederick Street,
Hornsby,
New South Wales 2077,
Australia

Renal endothelin and hypertension

SIR — Kurihara *et al.*¹, in their paper on the production of mice deficient for the gene for endothelin-1 (ET-1), report one result of great interest to those in the field of human hypertension. In their heterozygote ET-1^{+/-} mice, reductions in plasma and lung tissue levels of the mature peptide were confirmed, but this was accompanied by increased blood pressure. This hypertension was observed in both conscious and anaesthetized mice, and was not associated with any change in responsiveness to exogenous ET-1, or to inhibition of nitric oxide synthase. Some of the vascular mechanisms by which deficiency of ET-1 during ontogeny could cause hypertension have been discussed by Kurihara *et al.*¹ and by Vanhoutte in an accompanying News and Views article².

We would like to suggest a different mechanism based on the emerging role of the renal ET-1 system in blood pressure regulation. Kitamura *et al.*³ reported reduced tissue levels of ET-1 in kidneys of spontaneously hypertensive rats. Others⁴ have confirmed this observation. We recently reported markedly reduced urinary levels of ET-1 in patients with essential hypertension, especially in a subgroup whose blood pressure was sensitive to salt loading⁵. Decreased urinary levels of ET-1 are found in women with either pre-eclampsia or hypertension during pregnancy⁶. Because ET-1 is synthesized in the kidney, excreted in the urine, and not filtered from the plasma⁷, low urinary ET-1 could reflect a reduced rate of renal synthesis of ET-1.

The common denominator of all these studies, in both experimental animals and in humans, is an association between reduced renal ET-1 generation and elevated blood pressure. Because the accumulating evidence suggests that the physiological autocrine/paracrine function of ET-1 in the kidney is natriuresis and diuresis⁸, it is conceivable that renal deficiency of the peptide causes sodium and water retention and salt sensitivity, and thereby volume-overload hypertension.

These observations suggest that either an absolute or relative deficiency of renal

ET-1 may be a causative mechanism for the development and/or maintenance of human essential hypertension. Further studies should clarify the role of the renal ET-1 system in the physiological regulation of blood pressure, to identify those patients with hypertension associated with low urinary ET-1, and to explore specific diagnostic and therapeutic options.

Aaron Hoffman

Transplant Unit,
Rambam Medical Center and
Faculty of Medicine,
Technion, Haifa, Israel 31096

Ehud Grossman

Department of Medicine,
Sheba Medical Center and
Faculty of Medicine,
University of Tel-Aviv, Israel 52621

Zaid A. Abassi, Harry R. Keiser

Hypertension-Endocrine Branch,
National Heart, Lung and Blood Institute,
NIH, Bethesda, Maryland 20892, USA

MAEMURA *ET AL.* REPLY — Hoffman *et al.* suggest that elevated blood pressure in ET-1^{+/-} heterozygous mice¹ may be caused by the impairment of sodium and water handling. To test this hypothesis, we fed 8–12-week-old male heterozygous and wild-type mice with a normal chow con-

associated with increased sodium retention and volume expansion. Thus, we believe that factors other than volume overload due to sodium retention are responsible for blood pressure elevation in ET-1^{+/-} mice.

These results do not exclude a physiological role of ET-1 in renal water and sodium handling for the following reasons. First, the diuretic and natriuretic effects of ET-1 could be negated by an opposite effect on renal fluid absorption¹⁰. Second, the presence of impaired renal function and salt-sensitivity in ET-1^{+/-} mice can be manifested only by experiments using high salt loading¹¹. In young Dahl salt-sensitive rats, for example, a low salt diet does not cause sodium retention, volume overload and high blood pressure, but a high salt diet does¹². Third, any local effect in organs such as the kidney might be obscured in ET-1-deficient mice because their phenotype represents a net effect of systemic decrease in ET-1 production. Furthermore, a decrease in diuresis and natriuresis, if it occurs, may be compensated by other systems such as the renin-angiotensin system through possible feedback mechanisms in ET-1^{+/-} mice. Such a problem is a general limita-

COMPARISON OF RENAL PARAMETERS BETWEEN ET-1^{+/-} HETEROZYGOUS AND ET-1^{+/+} WILD-TYPE MICE

	Wild-type	Heterozygote
Renal ET-1 content (pg per mg protein)	2.3 ± 0.4 (n=13)	1.2 ± 0.1* (n=16)
Blood urea nitrogen (mg per dl)	26.6 ± 1.7 (n=12)	27.3 ± 1.0 (n=14)
Serum creatinine (mEq per l)	0.30 ± 0.03 (n=19)	0.29 ± 0.02 (n=22)
Serum Na (mEq per l)	150 ± 1 (n=12)	151 ± 1 (n=10)
uNa/uCr (mEq Na per mg creatinine)	0.24 ± 0.03 (n=6)	0.24 ± 0.02 (n=6)
Plasma volume (μl per g body weight)	51.8 ± 2.5 (n=7)	51.4 ± 1.6 (n=12)

uNa/uCr, urinary sodium excretion divided by urinary creatinine excretion; *P<0.05 versus wild type. Data are presented as mean ± s.e.m.

taining 0.2% NaCl in metabolic cages and measured several parameters relating to renal function and volume homeostasis. Under these conditions, the blood pressure of ET-1^{+/-} mice was significantly higher than that of wild-type mice¹. Circulating plasma volume was also determined using ¹²⁵I-labelled albumin⁹. In ET-1^{+/-} mice, kidney ET-1 levels were half those in ET-1^{+/+} wild-type mice (see table). There was no difference between heterozygous and wild-type mice in the other parameters examined (table). In particular, the absence of any difference in serum sodium concentration, urine sodium output or circulating plasma volume argues against the idea that the elevated blood pressure in ET-1^{+/-} mice is

tion on the interpretation of the physiological consequences in transgenic mice in which vasoactive substances have been 'knocked out'. We agree that further examination of physiological changes under various conditions is necessary to elucidate the precise mechanism of blood pressure elevation in ET-1^{+/-} mice and to clarify the role of ET-1 in various organs.

Koji Maemura

Yukiko Kurihara

Hiroyuki Morita

Hiroki Kurihara

Yoshio Yazaki

Third Department of Internal Medicine,
Faculty of Medicine, University of Tokyo,
Hongo, Bunkyo-ku,
Tokyo 113, Japan

- Kurihara K *et al.* *Nature* **368**, 703–710 (1994).
- Vanhoutte, P. M. *Nature* **368**, 693–694 (1994).
- Kitamura, K. *et al.* *Biochem. biophys. Res. Commun.* **162**, 38–44 (1989).
- Hughes, A. K. *et al.* *Hypertension* **20**, 666–673 (1992).
- Hoffman, A. *et al.* *Kidney Int.* **45**, 556–560 (1994).
- Wang, M. *et al.* *Am. J. Hypertension* **7**, 308–313 (1994).
- Abassi, Z. A. *et al.* *Hypertension* **20**, 89–95 (1992).
- Hoffman, A. *et al.* *Eur. J. Pharmacol.* **182**, 603–606 (1990).
- Crispell, K. R., Porter, B. & Nieser, R. T. *J. clin. Invest.* **29**, 513–516 (1950).
- Garcia, N. H. & Garvin, G. L. *J. clin. Invest.* **93**, 2572–2577 (1994).
- Campese, V. M. *Hypertension* **23**, 531–550 (1994).
- Simchon, S., Manger, W. M. & Brown, T. W. *Hypertension* **17**, 1063–1071 (1991).

The rise and fall of science

A. Rupert Hall

The Scientific Revolution. By H Floris Cohen. *University of Chicago Press: 1994.* Pp. 662. \$86.25, £59.95 (hbk); \$30, £21.50 (pbk).

BETWEEN 1934 and 1961, A. J. Toynbee published a 12-volume review of the rise and fall of civilizations entitled *A Study of History*. Professor Cohen's lesser, but still massive, book might have been entitled *A Study of the History of Science*, for it ranges over the rise and fall of the sciences from the Greeks to the nineteenth century, with substantial sections on Islam and China (the latter devoted to the great work of Joseph Needham). No such critical study of the writings of historians of science from William Whewell (1837) to 1985 has been published before, nor is it likely to be soon repeated. No other author has attempted so rich a survey of historians' views on the course of science in different cultures, comparing failures with the brilliant, seemingly endless expansion of knowledge of nature that has developed in the West since the seventeenth century.

This 'scientific revolution', as it has been labelled for 300 years (and Europe's sense of its progressive capacity is at least a century older still), culminating in the work of Isaac Newton, forms the core of Cohen's discussion — to use a favourite word of his. He is not interested, however, in any writers before Whewell, omitting also the many learned mathematicians and scientists who, in Whewell's lifetime and later, created the history of science as a scholarly subject and whose writings are still admired by contemporary scholars. Cohen's dictum that "To the extent that the history of scientific thought was cultivated at all, it was almost exclusively the province of the philosopher" can be accepted only if 'scientific thought' is interpreted as 'speculative thought'. The pioneers of the history of science (other than Whewell) did not, indeed, construct 'overviews' of its development, whereas Cohen's search among historians is for "generalisations [that provide] an enhanced understanding of how this peculiar [Western] civilization could bring forth the Scientific Revolution".

Cohen traces our present century's extensive historiography of the scientific revolution to the rejection, during its early decades, of the strong claim by Pierre Duhem (a physical chemist!) that the origins of modern science were to be found in fourteenth-century Paris, rather than in Poland or Italy at least two centuries later. Of this rejection Cohen gives an excellent account, attaching just importance to the contributions of Anneliese Maier, E. J. Dijksterhuis and Alexandre Koyré — who rightly remains an heroic, if

not impeccable, figure. The broadening of the idea from an originally narrow concept of a violent upheaval of new ideas and methods in astronomy and mechanics into an analysis of profound changes in the methods, ambitions, principles and doctrines of all the natural sciences is well handled. Yet Cohen hardly brings out the way in which the massive evidence of change in being — from the days of

extension of Darwinian evolution to the level of human society". Spencer's *Social Status* was published in 1850, nine years before Darwin's *Origin*.

Changes in and criticisms of the mid-century historians' concept of the scientific revolution are well presented. Cohen receives favourably the attempts of T. S. Kuhn and R. S. Westfall to impose a loose taxonomic structure on the multifarious activities of the "century of genius" (A. N. Whitehead's phrase), and due attention is given to religious causation (Hooykaas and others) as well as to the sociological analyses of Boris Hessen and Robert K. Merton; the strategic principles enunciated by Max Weber arouse his great admiration. But, surely with justice, Cohen will have no one-legged inter-



"There is not a Signor or great person who comes to Naples from far and remote parts who does not wish to see it out of curiosity and who is not overcome with amazement, after having seen it," boasted Francesco Imperato of the museum of nature created by his father, the apothecary Ferrante Imperato, and considered to be one of the wonders of the sixteenth and seventeenth centuries. This picture is taken from Ferrante Imperato's *Dell'istoria naturale* (1672), and is reproduced from *Possessing Nature* by Paula Findlen, a scholarly, detailed study of museums, collecting and scientific culture in early modern Italy. University of California Press, \$55.

Bacon, Gilbert and Harvey; of Fabricius and Libavius; of Gassendi and Descartes onwards — really makes the confinement of 'scientific revolution' to the mathematical sciences impossible. Nor, at the time, were events perceived in so limited a manner. Because Cohen does not pose such questions as "How did *this* historian think Galileo's genius operated?" or "Does *that* historian fairly understand and weigh the respective contributions to mechanics of Huygens and Newton?", the hard edge of scholarship tends not to be felt; Cohen is himself sometimes insensitive to it, as when he writes of "Herbert Spencer's

pretation. While the Platonic revival, so rewarding to a few seventeenth-century minds, is not emphasized, he sympathizes in a rather woolly way with those from Frances Yates onwards who believe that magic, hermeticism and alchemy, if not the 'onlie begetters' of modern science, at least nurtured it. These esoteric subjects, he suggests, "possessed insight into the endlessly complex makeup of the human personality", thus mitigating the harshness of the philosophy of matter-in-motion. So science benefited from a brush with the "wisdom literature" running through the world's philosophies and religions". Not

surprisingly, Cohen is also attracted by Needham's 'organicism'.

To quote Whitehead again (although *Science and the Modern World* (1926) is regrettably not among Cohen's authorities), "The various incidents which marked the rise of science [or factors aiding its prosperity —] the growth of wealth and leisure; the expansion of universities; the invention of printing; the taking of Constantinople; Copernicus; Vasco da Gama; Columbus; the telescope" — are mostly examined at length as possible causes of the scientific revolution. One, the European university, is wholly omitted. This is a pity, because the Western university has been a unique feature of our culture. Its association with the rise of modern science is clear, for all Duhem's exaggerations, and the writings of Hastings Rashdall, C. H. Haskins and their successors might well have been pondered by our author.

Though there is much in his historiography to question as well as to enjoy, though one may find the long comparison of China with Europe developed to little profit (since it leads, essentially, to the truism that East was East, and West was West, and rarely the twain did meet), and though one may regret the omission of the writings of very many good historians (like Willy Hartner, a scholar of universal capacity), Cohen's is a most valuable introduction to large questions involved in twentieth-century writing of the history of science. Its acute criticisms and creative suggestions make it more than that. How one wishes that anything comparable had been in print half a century ago! Today's reader with ample patience — for the author loves to take his time, to ruminate and to recapitulate — can acquire both information and perspective from Cohen, whose command of English is truly remarkable. □

A. Rupert Hall is at 14 Ball Lane, Tackley, Oxfordshire OX5 3AG, UK.

New editions

Nature's Economy: A History of Ecological Ideas by Donald Worster (2nd edn). Includes portraits of Linnaeus, Gilbert White, Darwin, Thoreau and twentieth century ecologists like Rachel Carson, Frederic Clements, Aldo Leopold, James Lovelock and Eugene Odum. Cambridge University Press, £40, \$59.95 (hbk), £13.95, \$14.95 (pbk).

Invention and Evolution: Design in Nature and Engineering by Michael French (2nd edn). Although intended primarily as a textbook for students of engineering, design and manufacturing technology, the book should also be of interest to professional engineers, designers and biologists. CUP, £50 (hbk), £17.95 (pbk).

Dispensing a dose of common sense

Alison Abbott

Forbidden Drugs: Understanding Drugs and Why People Take Them. By Philip Robson. Oxford University Press: 1994. Pp. 245. £9.99, \$16.95 (pbk).

A QUOTE from Groucho Marx opens Philip Robson's admirable attempt to explain drugs sanely to the layman: "I was teetotal until prohibition".

Robson himself confesses in his preface that he does his best "never to talk about drugs at social gatherings". Like Groucho, he believes that the illegal status of recreational drugs is irrational. But in his experience this intellectual position triggers a wild emotional response considered to be impolite in the British cheese-and-wine environment. This type of reaction also finds its place at the level of governments. In the United States there is a fervour to the (generally uninformed) debate on drugs that recalls McCarthyism.

Robson, a clinical psychiatrist from Oxford who has considerable practical experience with drug users, does not here advocate that all drugs should be legalized immediately. He argues instead that steps should be taken to wrest control of the drugs industry from organized crime, and that this inevitably requires some rewriting of drug laws.

Ignorance and half-truths fuel the fires of irrationality and Robson's aim is to bring a degree of sophistication to the debate by providing protagonists and opponents of legalization with the hard facts to support their arguments. The result is a highly readable, often entertaining, educational account of exactly what we know about recreational drugs: who uses them, how each class of drug works and the biological basis of addiction.

Only in the last chapter does Robson turn to his argument that drug laws should be reconsidered, his main argument being that the "war on drugs" has proved to be frighteningly expensive yet apparently ineffective. Heroin and cocaine are just as easy, if not easier, to buy on the street now as they were 15 years ago, and supplies of both are now cheaper and purer. Drug-related death has increased (by 40 per cent in Australia between 1981 and 1987), and more users are choosing to inject their drugs (AIDS transmission through shared needles is a sad consequence of this practice). Organized crime is the main beneficiary of this flourishing market, violence and petty crime the consequence that ordinary people have to suffer. The United States has the Western world's biggest drugs problem and one of the most repressive drug policies: "cause

or effect?" asks Robson.

That people fear legalization is understandable. Nearly all recreational drugs are poisonous (cannabis could be defined as an exception) and many are also highly addictive: witness alcohol. A legal supply through the state would control drug purity (we know the alcohol content of our wine and whisky) and at least help reduce the likelihood of unintentional poisoning. Addiction is harder to address, beyond noting that prohibition has had no effect on its incidence, but simply ensures that it is accompanied by crime. Robson argues for a transfer of taxpayers' money away from the war on drugs, into programmes to alleviate the social conditions that foster drug use. He argues not for total legalization of drugs but for "a move further along the path of controlled availability of opiates and stimulants from doctors" and legalization of the relatively benign cannabis.

Apart from this bold attack on drug policy, most of the book is dedicated to describing the drugs themselves: cannabis, the stimulants, hallucinogens, inhalants, ecstasy, tranquillizers and opiates. Demystification is the name of the game. Robson explains, with an admirable mixture of simplicity and accuracy, what users might expect from the drug they are about to take, what exactly they will experience, and the relative (documented) dangers of each type of drug, in terms of addiction potential, or short- or long-term toxicity and — rather too briefly — how each works biologically. He also talks about how users might perceive the advantages or disadvantages of their chosen drug. Sniffing glue may give you a short and unpredictable high and leave you headachy, but it is very cheap and easily available for school children. On the other hand, cocaine offers yuppies — or their 1990s' equivalents — advantages of enhanced social skills and energy for concentrated work that make up for subsequent crashes.

The book's historical accounts of abused drugs make fascinating reading whatever your view on legalization. Did you know, for example, that Coca Cola (which originally contained cocaine) was marketed in the late nineteenth century as a "refreshing and stimulating alternative to alcohol"? Or that in the First World War, the classy London department store Harrods sold morphine and cocaine kits, complete with syringes and needles, labelled "A Useful Present for Friends on the Front"? Or that Hitler was said to be injecting himself with amphetamines and other drugs towards the end of the Second World War?

Robson also describes the various models of addiction that have been put forward. He rejects psychoanalytical models and analyses better-documented biological models that link positive reinforce-

ment (the driving force to return to the substance abused) with dopaminergic pathways in the nucleus accumbens, in the context of socio-cultural models of addiction: social deprivation promotes drug abuse.

Politics and organized crime being what they are, it will of course take a lot more than this book seriously to influence the entrenched resistance to changing drug policy. This is a shame. Addictus, Robson reminds us, was a victim of the courts in ancient Rome, which handed him over as a slave to his creditor because he could not pay off his debts. A neat metaphor for Robson's views. □

Alison Abbott is Senior European Correspondent for Nature, based in Munich.

Thinking on one's feet

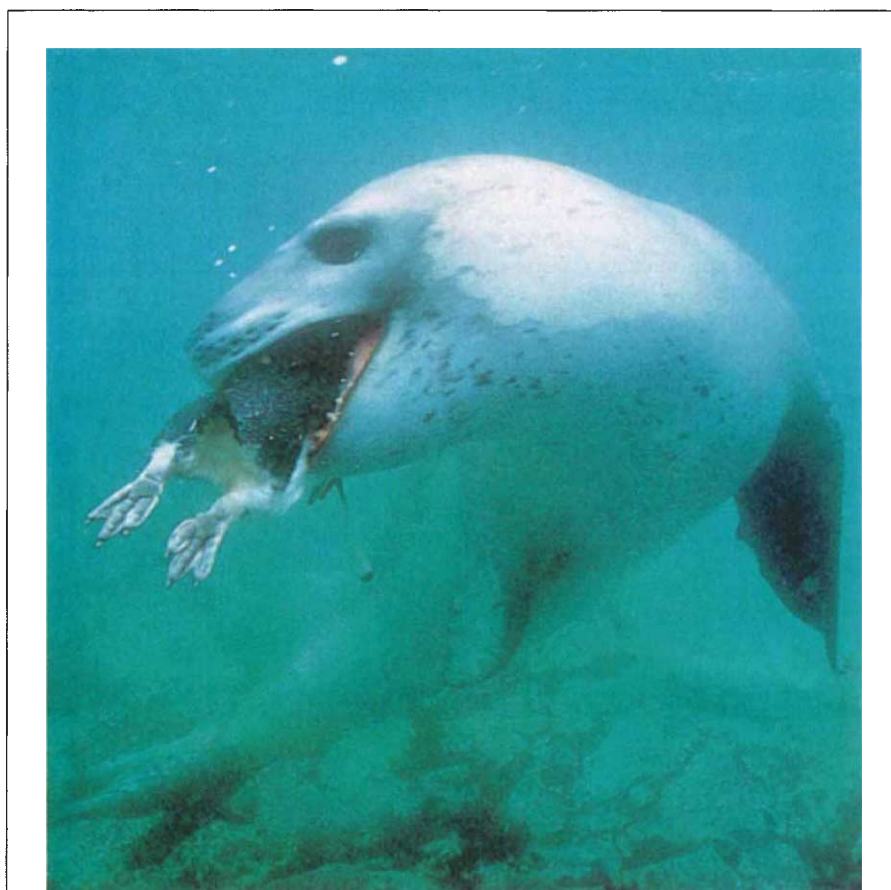
Annette Karmiloff-Smith and
Mark H. Johnson

A Dynamic Systems Approach to the Development of Cognition and Action. By Esther Thelen and Linda B. Smith. MIT Press: 1994. Pp. 376. £44.95, \$67.50.

DEVELOPMENTAL theories are going dynamic. Psychologists no longer think of nature and nurture in terms of a dichotomy. But until recently there were few alternatives to the biologically implausible dissociation of cognition into 'innate' and 'acquired' components. Two recent approaches — connectionist models of cognitive change and dynamic systems approaches to motor development — now promise to change the way many think about development.

This book attempts to set the agenda for broadening the dynamic systems approach to a wide range of phenomena in brain, cognitive and behavioural development. Thinking, according to Thelen and Smith, is like walking: the emergent property of multilevel dynamics. Irrespective of whether they succeed in their aims, this is a challenging book. The questions addressed go to the very heart of crucial developmental issues. How does experience shape the developing brain? Where do novelty and flexibility come from? How is knowledge assembled in real time? What is the role of temporal versus spatial contingencies? Do constraints on, say, language development imply a genetic blueprint for language or the emergence of constraints from a variety of different levels of interaction? Or in other words, what causes developmental change?

The dynamic systems approach holds that there is no single cause of developmental change. Although the endpoints of



A PENCHANT for penguins — a leopard seal *Hydrurga leptonyx* seizes a penguin underwater. Only males appear to indulge in this seasonal eating habit. Picture taken from *Seals and Sea Lions of the World* by Nigel Bonner (Blandford, £18.99).

human development are complex and unique, the processes by which we reach those endpoints are considered to be the same as those that govern development in very simple organisms, and even in some complex nonliving systems. Genetic determinism, the authors argue, simply sidesteps the problem of origins, dumping it onto the laps of evolutionists. Variable, task-sensitive effects are not just noise in a grand developmental plan, but are the raw material for developmental change. The central theoretical question is not how cognition reaches stability, but rather how cognition calibrates its continual stabilization—destabilization over real time and developmental time in the face of an ever-changing world.

The authors are to be congratulated on taking development seriously. Theirs is a radical departure from most of current cognitive development theory which focuses on innate specifications, symbolic representations and the attainment of stable structures. Many of the authors' arguments are reminiscent of those of Piaget. They give primacy to action, perception and movement as well as to the ability of the organism to modify and equilibrate itself. Where they differ from the Piagetian view is in arguing that

mental activity is a dynamic assembly of multiple levels (neural, social, physical) rather than a hierarchy of increasingly stable mental structures. The authors' dismissal of the need to specify the nature of mental representations, however, goes beyond disagreement with Piagetian theory.

The dynamic systems approach is readily applicable to behavioural output, as the book nicely illustrates. At this level, reducing the importance of the brain to only one of many sources of control parameters with the same status as muscle strength may be plausible. When we move to cognitive development, however, the special status of representations in the brain becomes essential, in our view. Not all developmental scientists consider the term 'representation' to mean something static. The authors rightly reject the digital computer metaphor for cognitive development popular in the 1970s and 1980s, but in denying the importance of representations, they throw the baby out with the considerable quantity of bathwater.

Their attempt to move from cyclical behavioural output to cognition (sans representations) manifests itself in two ways. First, they cite evidence from neuroscience that information processing in the

brain crucially depends on oscillatory processes, and, second, they reinterpret some well known experiments on the perception of objects by infants in terms of oscillatory perceptual input. In this way they reduce explanations in terms of representations to temporal dynamics. But beyond brain and perceptuo-motor behaviour, we believe it is essential to consider representational change and the formation of conceptual categories.

For example, Jean Mandler has demonstrated that infants do not rely only on their efficient abilities to detect perceptual similarity. They also build up conceptual categories. Given a plastic aeroplane and bird with outstretched wings that look extremely similar perceptually, infants nonetheless treat them very differently conceptually, one belonging to the category of mechanical objects, the other to biological beings. Conceptual categories can also be thought to emerge

from multiple levels of interaction of information, rather than innately specified distinctions, but they are still qualitatively different from the perceptual categories.

Nativists, structuralists, empiricists and social constructivists will disagree with different parts of this book. Yet this landmark volume is essential reading for all of them. The authors' reinterpretation of several well-known infancy experiments that are hotly debated by these various camps is real food for thought. The book might well make such readers question the presuppositions that sustain their current thinking and modify the way they conduct and interpret developmental research. □

Annette Karmiloff-Smith and Mark H. Johnson are in the MRC Cognitive Development Unit, 4 Taverton Street, London WC1H 0BT, UK.

Being clever and knowing it

Marian Stamp Dawkins

The Animal Mind. By James L. Gould and Carol Grant Gould. *W. H. Freeman/Scientific American Library*: 1994. Pp. 236. \$32.95, £19.95.

NOT so long ago — less than 20 years in fact — the subject of animal consciousness was seen to be not quite scientifically respectable, something to be touched on only by those who were either outside mainstream biology or at least old enough to have reached what Donald Griffin called the 'philosophical pause'. Real working scientists were supposed to stick to what animals could be objectively observed to do and to leave their minds severely alone. Since that time, however, the study of mind in both human and nonhuman animals has undergone something of a revolution. The study of human consciousness has not just become respectable scientifically but positively ultrafashionable, with books, conferences and a burgeoning number of different theories about what it might be.

It is therefore appropriate that James and Carol Gould have brought together what we know of the workings of the minds of nonhuman animals. Well written and beautifully (and informatively) illustrated, their book describes some of the 'cleverer' things animals have been claimed to do in a way that is calculated to arouse anyone's interest. It covers the way in which animal sense organs work as well as the most recent work that suggests that parrots may be able to 'count' and chimpanzees may be able to manipulate symbols in a language-like way. The book can, in fact, be recommended as a very read-

able introduction to the study of cognition in animals, with one major proviso.

The proviso is that the book blurs the important distinction between being clever and being conscious. It could leave the unwary reader with the impression that all we need to do to probe animal minds is to show that they are capable of various complex intellectual tasks and we will inevitably have shown that they are conscious. To give the book its due, the authors do not actually say this, but the fact that they fail to make a clear distinction between the issue of whether an animal (or a machine) can find a solution to an intellectual problem and whether that solution is found consciously could cause confusion. The phrase "cognitive finesse", for example, when applied to bees (p.113) has a high degree of ambiguity about it and could well be misunderstood.

There are many complex tasks that humans do without being conscious and we still do not understand the dividing line between what can be done unconsciously and what needs conscious awareness — or why. To decide on which side of this dividing line lie nonhuman animals, we certainly need facts about what animals are capable of doing, which the Goulds provide admirably. But we also need to be rather clearer than they are about the point at which we should take the further step of concluding that what animals do, they do consciously. □

Marian Stamp Dawkins is in the Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

New in paperback

The Elusive Transformation: Science, Technology, and the Evolution of International Politics by Eugene B. Skolnikoff. Princeton University Press, \$16.95. "Skolnikoff is talented at weaving the argument between science and technology and international developments...Full of stimulating and informative ideas" (Margaret Sharp, *Nature* **364**, 296; 1993).

Fossil Horses: Systematics, Paleobiology, and Evolution of the Family Equidae by Bruce J. MacFadden. Cambridge University Press, £15.95, \$29.95. "Timely and readable" (Adrian Lister, *Nature* **365**, 118; 1994).

Invention: The Care and Feeding of Ideas by Norbert Wiener. MIT Press, \$9.95, £8.95. For a review by Arnold Pacey see *Nature* **363**, 28 (1993).

Keywords in Evolutionary Biology edited by Evelyn Fox Keller and Elisabeth A. Lloyd. Harvard University Press, \$23.95, £15.95. For a review by J. S. Jones see *Nature* **363**, 220 (1993).

The Last Panda by George B. Schaller, with a new afterword. University of Chicago Press, \$15.95, £11.25. "Schaller lays out his recollections, emotions, wisdom and fears about pandas and their conservation. And he writes superbly — the book is quotably lively, varied and jumbled, riveting and often depressing, but essential reading (Brian Bertram, *Nature* **363**, 219; 1993).

Science and Anti-Science by Gerald Holton. Harvard University Press, \$17.95, £11.95. An "extraordinarily thoughtful, penetrating and wise tract for everyone concerned about the future of science" (John Ziman, *Nature* **367**, 522; 1994).

The Sexual Brain by Simon LeVay. MIT Press, \$10.95, £9.95. A look at the biological roots of human sexual behaviour. For a review by Steven Rose see *Nature* **363**, 505 (1993).

Uncovering the Past: A History of Archaeology by William H. Stiebing Jr. Oxford University Press, \$12.95.

What Makes Nature Tick? by Roger G. Newton. Harvard University Press, \$17.95, £11.95. For a review by N. David Mermin see *Nature* **366**, 724 (1994).

World Changes: Thomas Kuhn and the Nature of Science edited by Paul Horwich. MIT Press, £16.95. For a review by Peter Lipton see *Nature* **364**, 770 (1993).

Wrinkles in Time by George Smoot and Keay Davidson. Avon Books, \$12.50. For a review by Joseph Silk see *Nature* **366**, 373 (1993).

Mechanisms of intracellular protein transport

James E. Rothman

Recent advances have uncovered the general protein apparatus used by all eukaryotes for intracellular transport, including secretion and endocytosis, and for triggered exocytosis of hormones and neurotransmitters. Membranes are shaped into vesicles by cytoplasmic coats which then dissociate upon GTP hydrolysis. Both vesicles and their acceptor membranes carry targeting proteins which interact specifically to initiate docking. A general apparatus then assembles at the docking site and fuses the vesicle with its target.

THE central problem of modern cell biology is to understand the three-dimensional organization of cells in terms of underlying chemistry, usually involving protein-protein interactions. Such a topological biochemistry is needed to explain intracellular transport, the complex pattern of macromolecular traffic among organelles. The extensive network of intracellular membrane-bound organelles allows the eukaryotic cell to carry out a variety of specialized tasks, and greatly increases its surface-to-volume ratio. The existence of this network, however, raises the question of how the proteins characteristic of each compartment are conveyed to their destination (Fig. 1). This transport is necessary for the biogenesis of plasma membranes, lysosomes and endosomes, the secretion of proteins and other molecules from the cell, and the uptake of external molecules by endocytosis. Moreover, its specificity is harnessed to generate the distinct apical and basal surfaces needed for the polarized function of cells in most tissues.

The outline of the secretory pathway was delineated by Palade and his colleagues thirty years ago¹. Nascent proteins are delivered to the lumen of the endoplasmic reticulum (ER), pass through the Golgi complex for post-translational processing, and are stored in specialized secretory vesicles which fuse with the cell surface upon receipt of a signal for exocytosis. Movement of protein between these compartments occurs by the budding and fusion of transporting vesicles^{1,2}.

Subsequent studies from many laboratories extended this concept to other membrane systems, revealing the central role of the Golgi stack³. Electron microscope immunocytochemistry was used to follow the progress of newly synthesized proteins and to localize the Golgi compartments in which various post-translational modifications (mainly glycosylations) take place⁴. Proteins of the plasma membrane, secretory storage vesicles and lysosomes, for example, were found to leave the ER and cross the Golgi stack together. It is only during exit from the last, *trans* compartment that the various precursors are separated according to destination⁵. And while this traffic proceeds, additional vesicle transport pathways operate in independent and occasionally overlapping flow patterns.

Can this traffic be explained by the standard principles of solution chemistry? The finding that it occurs in cell-free extracts⁶, even after the precise relationships between organelles within cells are destroyed, indicates that it can. Such extracts are uniquely useful for establishing molecular mechanisms, and also furnish assays that have been used for the identification and/or purification of much of the required protein machinery. When combined with electron microscope immunocytochemistry, the structures they form can be readily related to their function. Cell-free systems reconstituting virtually every avenue of biosynthetic and endocytic protein transport in animal and yeast cells have now been described⁷.

The isolation of secretion mutants in yeast⁸ established that all eukaryotes have similar pathways, and offered a powerful approach to identify required polypeptides and test their physiological relevance. By collecting temperature-sensitive and other

secretion (*sec*) mutants, cloning and sequencing the affected genes, and analysing their phenotypes by electron microscopy, a considerable repertoire of transport machinery has been identified⁹ and its order of action in the protein transport pathway uncovered¹⁰.

The cell-free biochemical approach established the principles and molecular mechanisms underlying protein transport, and the genetic approach has contributed by establishing the *in vivo* relevance and the overall generality of these concepts. The story first emerged for intercisternal transport in the Golgi stack, and is still most complete there. This review therefore uses the Golgi as a paradigm before generalizing to other steps, and emphasizes transport pathways through the ER and Golgi, rather than other routes which are still less well understood. It will focus on the machinery that forms and fuses vesicles, rather than on the signals that determine whether individual proteins enter or avoid transporting vesicles at given steps, and will avoid explicit consideration of the methods involved except where advances have provided important information.

Transport across the Golgi stack *in vitro*

The Golgi stack is specialized for protein processing and transport, and vesicles with the surface area of an entire cisterna pass between one cisterna and the next every few minutes.

Transport between cisternae can be reconstituted by incubating Golgi membranes with cytosol and a source of energy such as ATP^{6,11}. When a 'donor' population of Golgi containing the G protein of vesicular stomatitis virus (VSV) is incubated with an 'acceptor' population of Golgi, containing an enzyme that adds *N*-acetylglucosamine, transfer of G protein from donor to acceptor can be measured by the addition of radioactive *N*-acetylglucosamine to the G protein¹¹. Transport between the successive Golgi compartments is unidirectional^{12,13} and as the glycosylated G protein is confined to the acceptor stack, which remains distinct from the donor stack¹⁴, it must involve diffusible intermediates like vesicles.

Electron microscopy suggested that the intermediates are indeed vesicles with a diameter of about 70 nm (ref. 15), which contain the VSV G protein and appear only when cytosol and ATP are added¹⁶. Most of the fully formed vesicles, and all of the vesicles still in the process of budding, have a distinct coat, which is ~18 nm thick, on their cytoplasmic surface (ref. 16 and Fig. 5a) and which resembles the clathrin coat of endocytic vesicles¹⁷. The coat is confined to budding regions and apparently forms only when Golgi membranes are incubated with cytosol¹⁶, indicating that the assembly of the coat from cytosolic precursors may drive budding.

Selective inhibitors reveal distinct stages in the maturation of these vesicles and help establish them as bona fide transport intermediates. Transport is blocked by non-hydrolysable analogues of GTP like GTP- γ S (ref. 18), and by low concentrations of the cysteine-alkylating agent *N*-ethylmaleimide (NEM). Electron microscopy shows that both treatments lead to the accumulation of 70-nm vesicles containing VSV G protein;

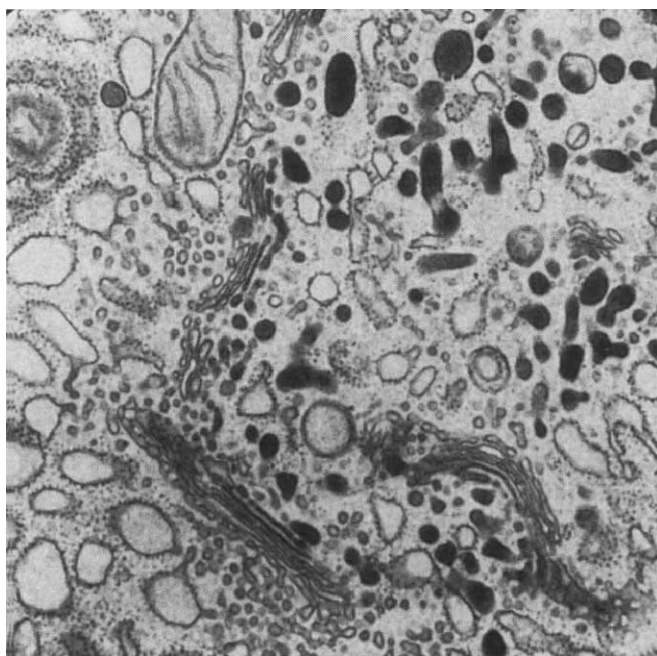


FIG. 1 The interface of endoplasmic reticulum and the Golgi stack, illustrating the structures, inter-relationships, and complexities of this region. Shown are ribosome-studded rough ER, ribosome-free transitional ER regions, with budding transport vesicles, the stack of Golgi cisternae with its distinct entry (*cis*) and exit (*trans*) faces, and the transport vesicles that surround the Golgi, especially in the region between transitional ER and the *cis* face of the Golgi stack with its associated tubular elements. These transport vesicles are mainly uncoated, most likely because the coat is removed after budding, and fusion is slower than budding. Specialized secretory cells such as this pancreatic exocrine cell (secreting digestive enzymes) store secretions in secretory vesicles that form (with other types of vesicles) from the *trans* face of the Golgi stack. Concentration of secretory proteins is already seen from the increase in luminal electron density in cisternae and vesicles on the *trans* side of the stacks. The secretory vesicles discharge their content upon fusion with the apical plasma membrane when a signal for exocytosis is received. From pancreatic lobules stimulated *in vivo* with carbamylcholine; this hyperstimulation results in the irregularly shaped, dense secretory granules, and in the accumulation of secretory products in *trans* Golgi elements. Magnification, $\times 18,800$. Courtesy of G. E. Palade.

GTP- γ S produces coated vesicles¹⁸, however, whereas NEM produces uncoated ones¹⁹. NEM and GTP- γ S together produce only coated vesicles²⁰, showing that the GTP- γ S block precedes the NEM block and that coated vesicles become uncoated after budding.

Transport across the Golgi thus appears to employ these coated vesicles, which are now termed COP (for coat protein)-coated vesicles. The coat must be removed after budding to allow fusion of the enclosed vesicle. The simultaneous inhibition of transport and uncoating by GTP- γ S indicated that uncoating may be driven by hydrolysis of GTP. As uncoated vesicles accumulate on Golgi cisternae, vesicle consumption by membrane fusion must require an NEM-sensitive protein. Identification of the proteins affected by GTP- γ S and NEM has indeed led directly to the heart of the mechanisms for budding and fusion.

Coatomer and ARF coat proteins

Purification of the vesicles that accumulate following treatment with GTP- γ S (ref. 21) shows that the coat contains eight polypeptides. One is ADP-ribosylation factor (ARF)²², a GTP-binding protein found mainly as a monomer in the cytosol²³. The other seven (termed COPs) are associated in an equimolar

coat protomer (coatomer) complex²⁴, which contributes most of the mass to the coat. Coatomer, like ARF, is mainly cytosolic, with only traces present in Golgi membrane fractions. ARF and coatomer therefore co-assemble on the Golgi surface to form vesicles; their later release constitutes uncoating. As Golgi membranes require only coatomer, ARF and nucleotides to form fusion-competent vesicles^{25,26}, other stoichiometric, cytoplasmic coat components are unlikely to be discovered.

Genetic studies with yeast have confirmed that coatomer, discovered and characterized *in vitro*, is essential for transport *in vivo*. Yeast has a coatomer whose γ -COP subunit is encoded by the *sec21* gene²⁷, needed for transport from ER to Golgi *in vivo*, and mammalian γ -COP is homologous to Sec21p (ref. 28). Microinjection of antibodies against β -COP has also been shown to block mammalian intracellular transport *in vivo*²⁹. The established role of COP-coated vesicles in *cis*-to-*trans* transport through the Golgi stack in no way precludes an additional role in retrograde transport from Golgi to ER³⁰.

ARF was first described as a cofactor required for ADP-ribosylation of a trimeric G protein α -chain by cholera toxin²³. It is concentrated in the Golgi, and yeast ARF is essential for Transport *in vivo*³¹. This is also true in mammalian cells^{32,33}. The finding that ARF is a stoichiometric component of COP-coated vesicles suggested that ARF functions in vesicle budding and is the target of GTP- γ S.

Coated vesicle budding and uncoating

The physicochemical properties of ARF and its presence in the coat suggested how GTP binding and hydrolysis might trigger budding and uncoating²². ARF is *N*-myristylated and its GDP-bound form is water-soluble; the GTP-bound form, however, is inserted into membranes in a myristic acid-dependent manner³⁴. The GDP-GTP conformational switch characteristic of G proteins thus appears in this case to modulate membrane insertion by controlling exposure of the covalently attached fatty acid.

Cytosolic ARF·GDP was proposed to be activated for budding by GTP-GDP exchange at the Golgi surface, producing Golgi-bound ARF·GTP, which would recruit coatomer from the cytosol to assemble the bud²² (Fig. 6, step 1). Subsequent hydrolysis of the bound GTP would induce uncoating by releasing ARF and coatomer (Fig. 6, step 4). If so, loading ARF with GTP- γ S should prevent uncoating, blocking transport and inducing an accumulation of coated vesicles, as observed.

Studies in the cell-free system have revealed this molecular mechanism to be essentially correct (Fig. 2). ARF binds to membranes in the absence of coatomer^{35,36}, and its binding must therefore initiate budding (Fig. 2a). Binding of soluble *N*-myristylated-ARF·GDP requires GTP and involves nucleotide exchange catalysed by an activity associated with Golgi membranes^{37,38}. Non-myristylated ARF undergoes nucleotide exchange but does not remain bound to the membrane. In addition to the ARF bound by its myristic acid moiety, Golgi membranes also carry a second population of ARF (presumably derived from the first) bound to an 'ARF receptor'³⁹. The molecular identity of the receptor is unknown, but its binding capacity is saturable, unlike that of the lipid membrane.

Saturable and specific binding of coatomer to Golgi membranes requires bound ARF, which cannot be replaced by any other cytosolic protein^{35,36}. The fact that coatomer binds to sites defined by ARF-ARF receptor complexes, and not to the free pool of ARF, implies that the ARF receptor (or associated polypeptides) must contact both ARF and coatomer for coat assembly.

The amount of coatomer bound at saturation (Fig. 2b) thus varies stoichiometrically with the amount of receptor-bound ARF^{35,36}. The resulting complexes consist of coated buds²⁶, virtually complete coated vesicles which are not yet pinched off at their base (Fig. 2c). The cisternal membranes remain flat on binding ARF²⁵, only forming buds when coatomer is added⁴⁰.



FIG. 2 Dissection of steps in the budding of COP-coated transport vesicles. *a*, ARF binds to Golgi membranes in the absence of other components upon activation by binding GTP, defining saturable ARF receptor sites, resistant to extraction with liposomes; taken from ref. 39. *b*, Coatamer binds to sites containing bound ARF to form stoichiometric complexes. The amount of coatamer bound at saturation (ordinate) depends on the amount of ARF bound to the Golgi membrane (abscissa); taken from ref. 36. *c*, The stoichiometric complexes consist of COP-coated buds (arrows), intermediates that must pinch off at their base to be released as coated vesicles; from ref. 26. *d*, The coated buds pinch off when incubated with palmitoyl-coenzyme A to yield fully formed coated transport vesicles²⁶, as shown here. Thin section shows the membrane of the vesicle surrounded by a dense outer coat of closely packed subunits. Magnification, $\times 89,250$. From ref. 21.

Coatamer binding and assembly thus drive budding, whereas ARF initiates budding by providing binding sites for coatamer. Most likely, membrane-bound ARF-coatamer complexes spontaneously assemble into arrays that form the coated bud. As expected, coatamer and ARF are both required to reconstitute vesicular transport⁴¹.

Even in the presence of excess coatamer, ARF and GTP, coated vesicles fail to pinch off. How is the nascent bud pinched

off, if not by coat assembly? Isolated Golgi membranes require fatty acyl-CoA as well as GTP, coatamer, and ARF to produce complete coated vesicles²⁶. Transport is stimulated by long-chain fatty acyl-CoA, such as palmitoyl-CoA⁴², and coated vesicle formation, like transport, is blocked by a non-hydrolysable analogue of palmitoyl-CoA⁴³. When Golgi stacks bearing coated buds are incubated with palmitoyl CoA, fully formed coated vesicles capable of fusion are released (Fig. 2*d*) in the absence of additional coat proteins²⁶.

Pinching off thus requires a process distinct from coat assembly (Fig. 6, step 3), involving 'periplasmic fusion' (that is, membrane fusion initiated from the luminal side of the bud, as opposed to vesicle fusion reactions, which are initiated from the cytoplasmic side). The process may thus resemble viral envelope fusion⁴⁴, perhaps involving a fusion protein activated by fatty acylation on the cytoplasmic side⁴⁵. The implications of this process for the topology and dynamics of intracellular membranes, including the formation of fenestrations⁴⁰ and tubular networks^{46,47}, are considered elsewhere⁵⁵.

The need for hydrolysis of ARF-bound GTP in uncoating was established by using a mutant version of ARF protein which binds GTP but cannot hydrolyse it⁴⁸. This mutant, like GTP- γ S, inhibits transport, with concomitant accumulation of coated transport vesicles. While budding still occurs, the resulting vesicles cannot be uncoated, and cannot fuse with acceptor Golgi²⁶, confirming that uncoating is a prerequisite for fusion. Uncoating therefore can be understood as a simple reversal of coat assembly: when ARF hydrolyses its bound GTP, its myristic acid will be 'retracted', ARF will dissociate from the membrane, and coatamer will follow suit. Additional (unknown) factors probably accelerate GTP hydrolysis by ARF to trigger uncoating and thereby initiate fusion.

Although membrane proteins mediating or controlling budding remain obscure (such as the ARF nucleotide exchanger and receptor, and trimeric GTP-binding proteins^{49,50}), a clear conceptual outline of the core process has emerged (Fig. 6, steps 1–4). A coat acts as a mechanical device to shape the membrane into a bud that can pinch off following periplasmic fusion at its base, and is released, when previously bound GTP is hydrolysed, to initiate fusion.

General fusion machinery

When transport is blocked by mild NEM treatment, uncoated vesicles accumulate on the acceptor Golgi cisternae^{19,20}, indicating that an NEM-sensitive fusion (NSF) protein must play a part in fusion (defined as any steps leading to content mixing once the transport vesicle binds to its target). As transport can be restored to such reactions by adding back fresh cytosol⁴², this fusion protein must be cytosolic, and the restoration of fusion provided an assay used to guide the purification of NSF⁵¹.

The slow intrinsic ATPase activity of NSF is necessary for fusion. Each of NSF's two homologous domains contains an ATP-binding site⁵², and mutation of either site seriously compromises both ATPase activity and fusion; mutating both sites eliminates ATPase activity and fusion⁵³.

The sequence of NSF⁵⁴ showed that it was homologous to the yeast protein Sec18p, previously shown to be somehow required for transport from ER to Golgi⁸. The discovery that NSF and Sec18p are equivalent⁵⁴ confirmed that the cell-free fusion pathway operates *in vivo* in cells ranging from yeast to mammals—a theme of universality that has persisted ever since. Indeed, even plant cytosol can fully replace animal cytosol in cell-free Golgi transport assays⁵⁵.

The equivalence of Sec18p and NSF also revealed that the NSF-dependent fusion pathway operates at several steps of transport, including ER-to-Golgi and within the Golgi stack. NSF or Sec18p are indeed necessary for every discernible step in transport from the ER to the plasma membrane and for endocytosis in animals and in yeast^{56–58}.

NSF itself does not bind to Golgi membranes until a crude cytosol fraction is added, providing an assay for the purification of the soluble NSF attachment proteins (SNAPs), which are also necessary for vesicle fusion^{59,60}. Of the three species of SNAP, α -SNAP and γ -SNAP have been found in all cell types examined so far. β -SNAP is confined to brain, closely resembles α -SNAP and is interchangeable with it in many assays. γ -SNAP is only weakly related to α/β -SNAPs⁶¹.

Sec17p is the yeast homologue of α -SNAP, as α -SNAP (but not β - or γ -SNAP) can restore the fusion activity of cytosolic extracts from yeast *Sec17* mutants⁶⁰. Electron microscopy independently established that Sec17p functions in fusion⁶². Sec17p and α -SNAP are indeed homologous, and Sec17p can mediate Sec18p binding to Golgi membranes⁶³.

SNAPs will bind to Golgi membranes in the absence of NSF, although the converse is not true, indicating that SNAPs bind before NSF. α - and β -SNAP bind competitively to the same site on alkali-extracted membranes, indicating that the SNAP receptor (SNARE) is an integral membrane protein. SNAP proteins do not interact with NSF in solution, but once bound to SNARE sites mediate saturable and stoichiometric NSF binding to Golgi membranes. γ -SNAP is not essential for NSF binding, but synergizes with α -SNAP to create an optimum binding site for NSF⁶⁴.

On binding to SNAP-SNARE complexes, NSF hydrolyses ATP and releases itself⁶⁴, but it can be trapped in its membrane-bound form by using the non-hydrolysable ATP analogue ATP- γ S. The complex of NSF, α -SNAP, γ -SNAP and the SNARE sites can be extracted intact from Golgi membranes with non-ionic detergent. This 'fusion particle' sediments at 20S, and may form the core of a general 'fusion machine'^{19,60,64}. The presence of SNAREs in this particle indicates that (like NSF and SNAP) they are important for fusion.

How can cytoplasmic proteins like NSF and the SNAPs participate in fusion? The simplest possibility is that they use the energy provided by ATP hydrolysis to create a 'fusion active' conformation which facilitates bilayer fusion.

A single polypeptide can catalyse membrane fusion, as shown by the viral envelope fusion proteins⁴⁴, and a small number of proteins may well mediate vesicle fusion once targeting is achieved. It is unclear whether NSF, SNAPs and SNAREs participate this directly in bilayer fusion, or whether these proteins alone are sufficient for fusion. SNAPs and NSF assemble on the SNARE complex, however, and hydrolysis of ATP by NSF is not only essential for fusion but also rearranges SNARE proteins⁶⁵. The biophysical mechanism of fusion remains a mystery.

Specificity in vesicle targeting

Maintaining the identity of membrane-bound compartments in the face of the massive flux between them requires a fusion mechanism of great specificity¹. A vesicle coming from the ER, for example, must fuse with the *cis* face of the Golgi stack rather than a lysosome or a mitochondrion if compartmentation is not to be lost within minutes. Although the general machinery described above can mediate fusion itself, the correct target must already have been chosen. How is this specificity achieved? Do docking proteins from vesicle and target membrane assemble together with SNAPs and NSF to form a fusion machine?

Because NSF and SNAP must promote fusion at many locations, there must be a family of SNAREs; each SNARE must be sufficiently distinct to direct it to a distinct location, yet sufficiently similar to allow the SNAPs and NSF to assemble on it. As the general fusion machinery should only be activated after specific vesicle docking, SNAREs may provide an important interface between vesicle docking and fusion, perhaps by participating in both processes.

Mammalian brain is a particularly convenient source of SNAREs⁶⁶, as it is highly specialized for the membrane fusion

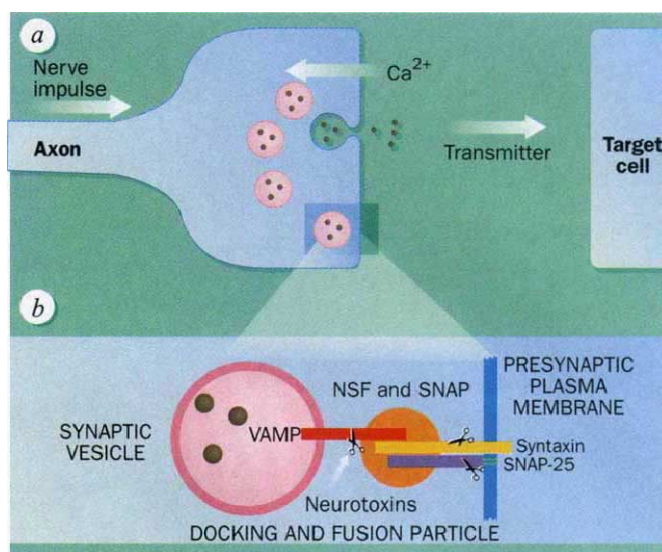
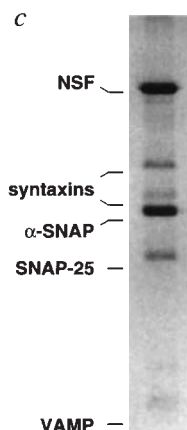


FIG. 3 Docking and fusion of synaptic vesicles. *a*, Synaptic vesicles cluster at the active zone and release neurotransmitter in response to a local, but prominent, depolarization-induced rise in cytoplasmic calcium. *b*, The 20S docking and fusion particle contains NSF and SNAP proteins (indicated by the orange disk), and VAMP, syntaxin and SNAP-25, which have been identified as select targets for particular neurotoxins causing botulism and tetanus⁷⁷. These toxins cleave the indicated target protein and block release of neurotransmitter, confirming the requirement for these subunits of the docking and fusion particle in synapses. The physiological role of NSF and SNAPs was confirmed by identifying corresponding secretion genes in yeast. *c*, Coomassie-blue-stained SDS-polyacrylamide gel illustrating the subunit composition of 20S particles purified from incubations of crude detergent extracts of brain membranes with recombinant NSF and SNAP. Reproduced from ref. 65.



involved in neurotransmission at synapses (Fig. 3*a*). Synaptic vesicles containing neurotransmitters may indeed account for more than 7% of the total brain protein⁶⁷. Moreover, the intensive study of synaptic transmission by molecular neurobiologists has uncovered the sequences and locations of most major synaptic vesicle proteins, as well as many proteins from the presynaptic plasma membrane with which these vesicles fuse^{67,68}.

To purify SNAREs, their biological properties were exploited in a two-stage affinity process⁶⁶. 20S particles were assembled from the SNAREs in a crude detergent extract of brain membranes by adding recombinant SNAPs and epitope-tagged NSF in the presence of ATP- γ S. After the tag had been used to effect a first purification step, ATP- γ S was replaced by ATP, allowing ATP hydrolysis and specifically releasing the SNAREs from NSF (the second biologically specific purification step). Three polypeptides were thereby isolated in virtually homogeneous form, released in a stoichiometric complex with the SNAPs. Sequencing revealed all of them to be synaptic proteins that had previously been cloned and localized but whose precise functions were unknown.

Two of these proteins, syntaxin (also known as HPC-1)^{69,70} and SNAP-25 (synaptosome-associated protein⁷¹ of M_r 25,000 (25K), not to be confused with the NSF-associated SNAPs), are localized to the presynaptic plasma membrane (Fig. 3). The other protein, VAMP for (vesicle-associated membrane protein,

also termed synaptobrevin), is confined to synaptic vesicles^{72,73}, and was already known to play a role in neurotransmitter release, as it is selectively cleaved by tetanus toxin, causing paralysis by blocking neurotransmission⁷⁴. The 20S particle involved in fusion of synaptic vesicles thus consists of NSF, SNAPs, VAMP, syntaxin and SNAP-25, without other stoichiometric components (Fig. 3*b* and *c*).

VAMP and syntaxin are both integral membrane proteins, with an amino-terminal cytoplasmic domain, a single presumed membrane-spanning segment, and a few mainly polar residues at their carboxy termini (Fig. 3). This topography seems ideally suited for docking (cytoplasmic domain) and possibly fusion (membrane-spanning domain); it is also characteristic of the viral envelope fusion proteins, although it is inverted to reflect the side of the membrane from which they initiate fusion⁴⁴. SNAP-25 is intrinsically water-soluble, but is anchored to the presynaptic plasma membrane, and behaves as an integral membrane protein, probably because it is fatty-acylated. Its four cysteine residues are closely clustered in the sequence and are likely acylation sites⁷⁵.

Synaptic 20S fusion particles contain both a protein derived from the synaptic vesicle (VAMP) and two proteins derived from the pre-synaptic membrane (syntaxin and SNAP-25), implying that the vesicle must dock when this complex forms⁶⁶ (Fig. 3*b*, *c*). These proteins (VAMP, syntaxin and SNAP-25) are confined to synapses and certain related tissues, but SNAP and NSF must interact with SNAREs in many subcellular membranes. Each membrane involved in an NSF-dependent fusion process would then be expected to have homologues of these synaptic proteins mediating docking and fusion together with the SNAPs and NSF⁶⁶. As the specificity of the system cannot be due to NSF and the SNAPs, it must reflect specific pairing between SNAREs in the vesicle and in the target membrane⁶⁶.

This suggests a hypothesis for vesicle targeting (the SNARE hypothesis)⁶⁶ in which each transport vesicle bears a unique address marker consisting of one or more v-SNAREs (generally VAMP homologues) obtained from its parental membrane during budding, while each target membrane is identified by one or more t-SNAREs (generally homologues of syntaxin and/or SNAP-25). Targeting specificity would thus be achieved by v-SNAREs binding to matching t-SNAREs (Fig. 4). If the interaction between cognate SNAREs were strong enough, docking could be achieved by SNAREs alone without additional binding energy provided by interactions with the SNAPs and NSF.

Evidence for the SNARE hypothesis

The SNARE hypothesis predicts⁶⁶ that all eukaryotic cells should have families of v- and t-SNAREs, whose members should form cognate pairs on the basis of binding specificity and should be essential for vesicle docking *in vivo* (Fig. 4*a*). Moreover, the overall program of vesicle traffic should be reflected in the intracellular localization of cognate SNAREs (Fig. 4*b*). The t-SNARE(s) cognate to the ER v-SNARE should thus be confined to the *cis* Golgi, whereas the t-SNAREs cognate to the *cis*-Golgi v-SNARE should be concentrated in the next Golgi compartment, and endocytic vesicles should contain yet other v-SNAREs which recognize t-SNAREs in endosomes, and so on. A mechanism should also exist to ensure that budding vesicles contain v-SNAREs but not t-SNAREs. Each SNARE should have signals directing it to the appropriate compartment, and a means to recycle transported v-SNAREs from acceptor compartments should also exist if the distinct chemistries of subcellular compartments are to be preserved⁷⁶. Speculative extensions of the hypothesis can explain phenomena such as retrograde transport, homotypic fusion and stacking of Golgi cisternae; these are considered elsewhere⁴⁵.

Many of these predictions have now been confirmed. Studies with tetanus and botulinum toxins have shown that the three synaptic SNAREs and SNAP are all essential for neurotrans-

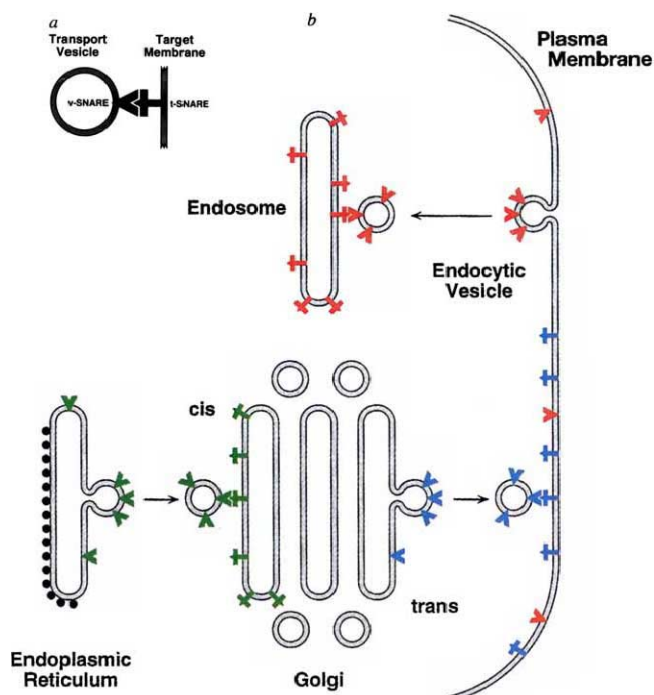


FIG. 4 The SNARE hypothesis. *a*, Each transport vesicle is proposed to be endowed with one or more unique v-SNAREs (generally related to VAMP) which pair with cognate t-SNAREs (generally related to syntaxin and/or SNAP-25), thereby docking the vesicle to the correct target. Other proteins may be involved in regulating the assembly of this core docking complex, which then serves as a scaffold for assembly of the general fusion machinery yielding a 20S docking and fusion particle. *b*, The road map of vesicular transport is proposed to be preordained in the pattern of localization of SNAREs among compartments. Only cognate v-SNAREs and t-SNAREs (Vs and Ts having the same colour) are detailed, corresponding to a few of the many vesicle shuttles known to occur in cells.

mission, as each synaptic SNARE is the unique target for cleavage by one or another species of toxin⁷⁷ (Fig. 3*b*). These results, together with the demonstration that NSF and SNAP are necessary for secretion in yeast, leave no doubt of the physiological relevance of the entire 20S particle (Fig. 3*b*).

It was independently noted^{78,79} that yeast cells contain proteins necessary for individual segments of the secretory pathway which are related to VAMP and syntaxin and are found in vesicle and target membranes, respectively (Table 1). The ER-derived v-SNAREs in yeast form a specific complex with the *cis*-Golgi t-SNARE which accumulates *in vivo* when fusion is blocked by inactivating NSF/Sec18p (ref. 80). Although these experiments provide striking biochemical and genetic confirmation of the SNARE hypothesis, they do not establish whether v- and t-SNAREs are directly paired within the 20S particle as the hypothesis proposes. Detailed studies of the synaptic proteins have been necessary to address this point.

Synaptic SNARE complexes are themselves sufficiently stable to allow their isolation. The yield of complexes containing stoichiometric amounts of VAMP, syntaxin and SNAP-25 from detergent extracts of brain membranes is unaltered whether they are free or assembled into 20S particles with SNAP and NSF⁶⁵. Other stoichiometric components are lacking from the free SNARE complexes, demonstrating a direct and specific inter-

TABLE 1 SNAREs, core compartment-specific docking machinery

	v-SNAREs (transport vesicles and donor compartments)	t-SNAREs (acceptor compartments)
ER	BOS1, BET1, SEC22	
cis-Golgi		SED5, Syn5
trans-Golgi	SNC1, SNC2	
Non-neuronal plasma membranes		SSO1, SSO2 Syn2, Syn4
Endosomes	Cellubrevin	
Presynaptic plasma membrane		Syntaxin SNAP-25
Synaptic vesicles	VAMP (synaptobrevin)	

Identity and intracellular localization of several known SNAREs. v-SNAREs are structural and/or sequence homologues of VAMP. t-SNAREs are homologues of syntaxin or SNAP-25. Components in lower-case are from animals; upper-case are from yeast, in which case mutation causes a secretory defect. Original references concerning yeast SNAREs are cited in ref. 78. Cellubrevin and Syn2-5 are described in refs 121 and 120, respectively.

action between VAMP, syntaxin and SNAP-25 alone, which therefore provides the core of the docking complex. Complexes between synaptic v- and t-SNAREs can then bind stoichiometric amounts of α -SNAP and NSF, suggesting that these fusion components bind after docking⁶⁵.

The cytoplasmic domains of syntaxin and VAMP bind each other alone⁸¹ but this complex is greatly stabilized by SNAP-25 (ref. 82). All three proteins can be predicted to participate in a coiled-coil structure, consisting of two or more strands of α -helices wrapped around each other. Whether a coiled-coil structure is actually formed is unknown, but the possibility that the specificity of membrane fusion is encoded in the zipping-up of cognate coiled-coil domains in vesicle and target membranes (as in transcription factors⁸³) is intriguing.

SNARE complex assembly

Mutations in the yeast *sec4* gene disrupt transport from the Golgi to the cell surface, leading to the accumulation of post-Golgi transport vesicles containing Sec4p (refs 8, 84). Sequencing of Sec4p (ref. 85) showed that it was a homologue of the mammalian GTPase, Ras. Soon thereafter, another yeast Ras homologue was identified (Ypt1p), a mutation in which disrupts ER and Golgi transport, and leads to vesicle accumulation⁸⁶. Antibodies against Ypt1p block vesicle docking and fusion *in vitro*^{87,88}.

The discovery of Sec4p and Ypt1p prompted the important hypothesis⁸⁹ that a family of Ras-related GTPases confers specificity on docking, using the GDP- and GTP-bound conformers of such proteins as a molecular switch registering whether the vesicle has docked.

Since then, a large family of GTP-binding proteins related to Ras (termed 'Rab' proteins) has been identified in various compartments of animal cells whose members are required at various steps of biosynthetic transport, endocytosis or recycling⁹⁰. All Rab proteins are intrinsically hydrophilic, like Ras itself, but are attached to membranes via C-terminal isoprenoid (geranyl-geranyl) and fatty acid chains.

Rab proteins clearly have a general role in vesicle docking, but their precise function is still obscure. The view that they are the principal source of docking specificity, however, now seems untenable for several reasons. A given Rab protein can be attached to more than one intracellular compartment and required for more than one transport step⁹¹. Moreover, a single hybrid of Ypt1p and Sec4p can provide Rab function to the

entire secretory pathway from the ER to the cell surface^{92,93}. Finally, the requirement for Ypt1p can be by-passed by over-expressing v-SNAREs⁹⁴. This indicates that although Rab proteins may normally be essential, conditions nevertheless exist in which Rab function is not needed for accurate vesicle transport and targeting.

What, then, is the role of Rab proteins? Because the Ypt1p is needed for the assembly of ER-Golgi SNARE complexes *in vivo*⁸⁰, Rab action must somehow facilitate the assembly of SNAREs. Yet it forms no part of the isolated docking complex⁸⁰. In a broad sense, Rab proteins must therefore catalyse v- and t-SNARE interaction. They may well participate in a complex reaction involving other proteins which either present the v-SNARE to the t-SNARE in a kinetically competent state or remove mispaired SNAREs in a proof-reading process.

Candidate proteins involved in this process include the compartmentally specific Sec1p protein family⁹⁵, whose neuronal homologue binds to a t-SNARE, thereby preventing the cognate v-SNARE from binding⁸². The precise role of Rab proteins may only be understood in relation to these and other proteins which are not yet integrated into a coherent framework. It makes sense, however, that a process as vital as docking should be closely regulated and possibly proof-read, and that there may be several layers to the targeting process.

Coupling of fusion to budding

A shuttling vesicle must be able to move molecules between two compartments while retaining the boundary that separates them, requiring that it completes its budding before it initiates fusion. As the v-SNAREs in vesicles must also be present in donor membranes, what prevents the donor and acceptor compartments from docking and fusing directly, without an intervening vesicle? A mechanism must clearly exist to ensure that fusion follows budding.

An important part of this device appears to be the coat, whose removal is required for the enclosed vesicle to fuse (Fig. 6, step 4). This was revealed⁹⁶ by the unexpected and dramatic effects of the natural product brefeldin A (BFA) on secretion and intracellular membrane organization (reviewed in ref. 46). BFA potently inhibits protein transport out of the ER and induces extensive absorption of the Golgi stack into the ER by membrane fusion. Soon after adding BFA *in vivo*, Golgi cisternae fuse to each other and actively extend along microtubules, forming membrane tubules that fuse with the ER. Tubulation is not an essential feature of the phenomenon, however, as it is stopped by disrupting the microtubules, which merely slows fusion of the Golgi with the ER. Even more remarkably, when BFA is removed, the Golgi apparatus reforms within a few minutes, indicating that the ER and Golgi have an intrinsic and still unexplained capacity to self-sort. This may be related to the unknown mechanisms by which v-SNAREs and the like return to donor compartments after fusion.

How does BFA trigger these effects? ARF and coatamer dissociate from the Golgi within seconds of BFA addition *in vivo*, before membrane fusion is induced⁹⁷. BFA does not actually disassemble coats, however, but blocks the binding of ARF and coatamer to the Golgi and thus the budding of coated vesicles⁹⁸ (Fig. 5a, b). Given the rapidity of vesicle budding and uncoating, BFA soon leads to the release of ARF and coatamer, and triggers a massive fusion of Golgi cisternae with each other (Fig. 5b). As ARF must be activated by nucleotide exchange to allow coatamer binding^{37,38}, BFA inhibits coat assembly primarily by preventing ARF binding (Fig. 6, step 1). This has recently been confirmed in whole cells³².

BFA's mechanism of action offers strong pharmacological evidence for the role of COP-coated vesicles in biosynthetic protein transport *in vivo*, complementing the immunological evidence²⁹ and the genetic evidence in yeast^{27,28}. It also implies that the coat not only drives vesicle budding but also prevents fusion prior to the completion of budding: when coat assembly is

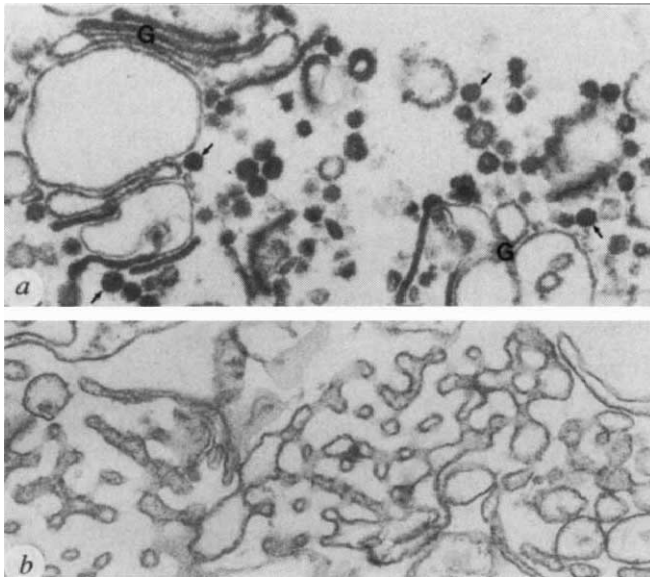


FIG. 5 a, Membrane pellet from a cell-free reaction illustrating vesicular transport between Golgi cisternae (G). Numerous fully formed or budding COP-coated vesicles (arrows) are present. Membrane fusion is tightly coupled to vesicle budding; that is, fusion of a vesicle does not begin before it is pinched off and its coat is removed. Magnification, $\times 65,000$. b, When coat assembly is prevented by adding brefeldin A (a compound that prevents GTP binding to ARF; Fig. 6, step 1), vesicle budding and vesicular transport stop, but membrane fusion continues, now uncoupled from vesicle budding. Fusion is no longer obliged to follow budding, and as a result the normally distinct cisternae of the Golgi stack rapidly fuse to yield a single compartment. Magnification, $\times 43,900$. From ref. 98.

stopped by adding BFA, vesicle transport stops and the involved compartments now fuse directly. BFA thus uncouples⁴¹ fusion from budding by preventing coat assembly.

How can the coat prevent fusion of the bulk of the compartment to which it is bound, given that membrane-bound coatomer and ARF are mainly located in coated buds, which only cover a small fraction of the Golgi membrane surface? Most v-SNARE molecules are incorporated into ER-derived vesicles during a single round of budding¹²⁶. Efficient sequestration of the v-SNAREs into budding vesicles (Fig. 6, step 1) would prevent engagement with t-SNAREs until the coat is removed following budding (Fig. 6, step 4), thus ensuring that fusion follows budding. When coat assembly is prevented (for example by BFA), free v-SNAREs will accumulate in the bulk of the donor membrane which can directly interact with t-SNAREs on the acceptor membrane, leading to the direct (uncoupled) fusion of donor and acceptor compartments.

The dual roles of coatomer and ARF as fusion inhibitor and budding activator have indeed been demonstrated in the cell-free Golgi transport system⁴¹. Removing ARF or coatomer from this system triggers uncoupled fusion^{41,99}, as does too low a concentration of cytosol (which makes ARF limiting and inhibits coat assembly)^{18,41,99}.

Generality of budding mechanisms

The principles established in the study of Golgi vesicular transport are rapidly being generalized as other transport steps are reconstituted and analysed in cell-free systems. It appears that a wide variety of vesicles may be shaped by coats whose assembly is triggered by activation of ARF (or a closely related protein) by GTP binding. ARF (but not coatomer) is required for budding of secretory vesicles and constitutive transport vesicles from the *trans*-Golgi network (TGN) (F. A. Barr and W. B. Huttner, manuscript submitted), which therefore seem likely to have a

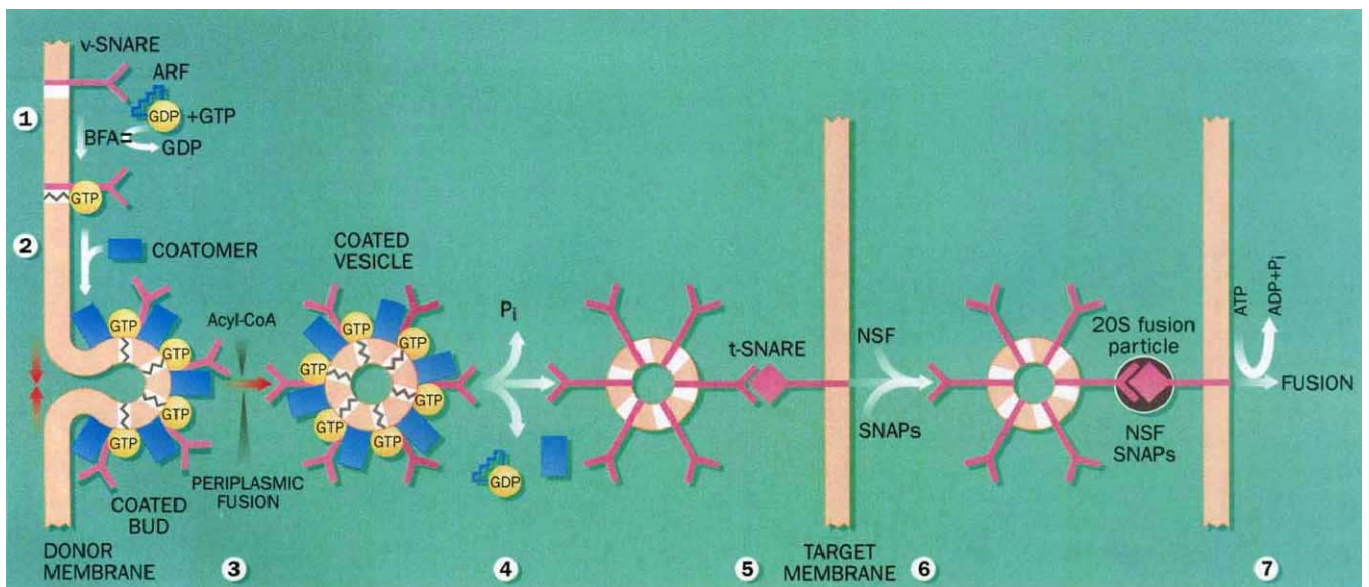


FIG. 6 Outline of steps in a round of transport. Coat assembly is initiated when a small GTP-binding protein (like ARF) is activated by binding GTP (step 1); the GTP-binding protein binds its receptor and recruits coat proteins (like coatomer), thus forming the coated bud (step 2). The bud pinches off in a distinct process involving periplasmic fusion (step 3) and the coat later disassembles following hydrolysis of the bound GTP (step 4). During budding, transport vesicles are tagged with members of a family of related integral membrane proteins, termed v-SNAREs, which serve as specific address markers. v-SNAREs pair with cognate t-SNAREs in the target membrane to dock the vesicle (step 5). The general fusion machinery then assembles on this scaffold (step 6). Much if not all of this machinery consists of cytoplasmic proteins that are

readily available to all cellular membranes, including the ATPase NSF and the SNAP proteins. SNAPs bind to the SNARE complex, enabling NSF to bind. ATP hydrolysis by NSF is essential for fusion (step 7), leading to bilayer fusion according to an unknown biophysical mechanism. The association between ARF receptor (white box, top left), whose molecular identity is unknown, and v-SNARE shown in step 1 is entirely speculative, but some means are needed to pack v-SNAREs into buds. The coatomer²⁴ consists of an equimolar complex of α -COP, β -COP^{122,123}, β' -COP^{124,125}, γ -COP, δ -COP, ϵ -COP, and ζ -COP. Other coat complexes are known, including the clathrin-adaptor coats¹⁷ and COPII coat¹⁰⁶.

non-COP coat that may correspond to the recently described 'lace-like' coat¹⁰⁰. Animal cells have several ARF proteins³⁴, which may reflect a still undiscovered multiplicity of coats. Only the role of ARF1 has so far been well studied. Some of these ARF species may be used in signal transduction pathways¹⁰¹, presumably unrelated to vesicle budding because the role of ARF in coat assembly is stoichiometric, not catalytic.

That coat assembly can drive budding was first suggested¹⁰² by electron microscopy of endocytic vesicles coated with clathrin¹⁰³, but the mechanism by which such vesicles bud is still unclear. Assembly of such vesicles at the TGN *en route* to endosomes is likely to involve clathrin binding to the Golgi, however, which requires activation of an ARF protein by GTP binding^{104,105}. This indicates that clathrin and COP-coated vesicles bud in related ways.

Studies with a cell-free system from yeast have led to the recent discovery of COPII-coated vesicles¹⁰⁶ which can mediate ER-to-Golgi transport *in vitro* and *in vivo*. The budding mechanism strikingly resembles that of COP-coated vesicles. Sar1p, a small GTP-binding protein related to ARF¹⁰⁷, plays a role analogous to that of ARF and a complex of other polypeptides (Sec23p/Sec24p and Sec13p/p150 (150K protein)) form the coat¹⁰⁶. Sec23p is a GTPase-activating protein (GAP) for Sar1p (ref. 108), and could readily ensure that uncoating follows budding since it only comes into contact with Sar1p·GTP once the coat is assembled, providing this complex lasts longer than the time required for periplasmic fusion.

In addition to the evidence above that COPII-coated vesicles mediate ER-to-Golgi transport there is also considerable genetic^{27,28,31}, biochemical^{29,109,110} and morphological¹¹¹ evidence that COP-coated vesicles also mediate ER-to-Golgi transport, in addition to their roles in transport within the Golgi stack. The simplest possibility is that COP and COPII-coated vesicles provide two parallel pathways from ER to Golgi, perhaps carrying different sets of cargo.

Triggered exocytosis

Much research on the discharge of secretory storage vesicles has concentrated on the cell-surface receptors and intracellular signalling systems that activate release by triggering vesicle fusion with the plasma membrane. The general apparatus underlying these pathways, which reflect the cell- and tissue-specific use of exocytosis (and are central to physiology), closely resembles that responsible for other forms of vesicle fusion.

The role of the NSF/SNAP-dependent pathway in both constitutively 'on' fusion events, such as intracellular transport, and regulated fusion events, like triggered exocytosis⁶⁶, which are kept 'off' in the absence of a special signal, raises the question of how the general fusion programme can be kept inactive until a signal for exocytosis is received.

One possibility is that one or a series of 'fusion clamps' stop progress along the general pathway at critical junctures^{65,66}. These clamps would be removed in response to second messengers inducing exocytosis, such as elevated cytoplasmic (Ca²⁺) concentrations. Such clamps would generally be membrane proteins so as not to stop other fusion processes (which must remain constitutive in most cells). How second messenger signals interface with NSF, SNAPs, SNAREs and other general fusion machinery is generally unclear, except in the case of synaptic transmission.

Here, the synaptic vesicle calcium binding protein synaptotagmin plays the primary role in triggering exocytosis when calcium is elevated following an action potential^{112,113}, although other triggering mechanisms also exist¹¹³. How could synaptotagmin clamp fusion? Synaptotagmin changes conformation when it binds calcium¹¹⁴. It also competes with α -SNAP for binding the synaptic SNARE complex⁶⁵, and could therefore prevent fusion by blocking assembly of the general fusion apparatus. Upon entry of calcium, synaptotagmin could dissociate from the SNAREs, allowing α -SNAP to bind and fusion to proceed⁶⁵.

Electrophysiological studies of exocytosis in single cells have shown that vesicle mobilization and fusion are kinetically complex, involving a series of calcium-dependent intermediates which eventually lead to the opening of a 'fusion pore', representing the moment of bilayer fusion¹¹⁵⁻¹¹⁷. These populations may reflect the position of clamps along the general fusion pathway, freezing vesicles at successive intermediate stages. Although many synaptic vesicles are docked at the presynaptic plasma membrane, only a few fuse in response to a given action potential; the rest are presumably less mature precursors. It is likely that the synaptotagmin clamp is applied immediately prior to fusion, downstream from this complex supply line of fresh vesicles.

Conclusions

About 35 years after the process of membrane fusion was first described¹¹⁸, a molecular basis has been established and transport between membrane-bound compartments can now be fitted into a coherent conceptual framework (Fig. 6). A similar machinery operates at multiple compartments in all eukaryotic cells. Exocytosis is regulated by superimposing inhibitory 'fusion clamps' on the general fusion pathway, which are removed in response to second messenger stimulation. Although many important details remain to be established, the overall pathway of budding, targeting and fusion is now clear.

Our prediction ten years ago that "All the roads for membrane traffic may yet prove to be paved with the same kinds of molecules"¹¹⁹ has undoubtedly been fulfilled. In the next ten years we can look forward to a richness of detail and mechanism broadening our understanding of the central themes, as the basis of cellular organization is further revealed in the wake of continuing advances in topological biochemistry. □

James E. Rothman is in the Cellular Biochemistry and Biophysics Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA.

- Palade, G. E. *Science* **189**, 347-358 (1975).
- Jamieson, J. D. & Palade, G. E. *J. Cell Biol.* **34**, 598-615 (1967).
- Farquhar, M. G. & Palade, G. E. *J. Cell Biol.* **91**, 77-103 (1981).
- Dunphy, W. G. & Rothman, J. E. *Cell* **42**, 13-21 (1985).
- Griffiths, G. & Simons, K. *Science* **234**, 438-444 (1986).
- Fries, E. & Rothman, J. E. *Proc. natn. Acad. Sci. U.S.A.* **77**, 3870-3874 (1980).
- Rothman, J. E. *Meth. Enzym.* **219** (1992).
- Novick, P., Field, C. & Schekman, R. *Cell* **21**, 205-215 (1980).
- Schekman, R. *Curr. Opin. Cell Biol.* **4**, 587-592 (1992).
- Novick, P., Ferro, S. & Schekman, R. *Cell* **25**, 461-469 (1981).
- Balch, W. E., Dunphy, W. G., Braell, W. A. & Rothman, J. E. *Cell* **39**, 405-416 (1984).
- Dunphy, W. G., Fries, E., Urbani, L. J. & Rothman, J. E. *Proc. natn. Acad. Sci. U.S.A.* **78**, 7453-7457 (1981).
- Rothman, J. E., Urbani, L. J. & Brands, R. J. *Cell Biol.* **99**, 260-271 (1984).
- Braell, W. A., Balch, W. E., Dobberty, D. C. & Rothman, J. E. *Cell* **39**, 511-524 (1984).
- Balch, W. E., Glick, B. S. & Rothman, J. E. *Cell* **39**, 525-536 (1984).
- Orci, L., Glick, B. S. & Rothman, J. E. *Cell* **46**, 171-184 (1986).
- Pearse, B. M. F. & Robinson, M. S. A. *Rev. Cell Biol.* **6**, 151-171 (1990).
- Melançon, P. et al. *Cell* **51**, 1053-1062 (1987).
- Malhotra, V., Orci, L., Glick, B. S., Block, M. R. & Rothman, J. E. *Cell* **54**, 221-227 (1988).
- Orci, L., Malhotra, V., Amherdt, M., Serafini, T. & Rothman, J. E. *Cell* **56**, 357-368 (1989).
- Malhotra, V., Serafini, T., Orci, L., Shepherd, J. C. & Rothman, J. E. *Cell* **58**, 329-336 (1989).
- Serafini, T. et al. *Cell* **67**, 239-253 (1991).
- Kahn, R. A. & Gilman, A. G. *J. Biol. Chem.* **261**, 7906-7911 (1986).
- Waters, M. G., Serafini, T. & Rothman, J. E. *Nature* **349**, 248-251 (1991).
- Orci, L., Palmer, D. J., Amherdt, M. & Rothman, J. E. *Nature* **364**, 732-734 (1993).
- Ostermann, J. et al. *Cell* **75**, 1015-1025 (1993).
- Hosobuchi, M., Kreis, T. & Schekman, R. *Nature* **360**, 603-605 (1992).
- Stenbeck, G. et al. *FEBS Lett.* **314**, 195-198 (1992).
- Pepperkok, R. et al. *Cell* **74**, 1-12 (1993).
- Cosson, P. & Letourneur, F. *Science* **263**, 1629-1631 (1994).
- Stearns, T., Willingham, M. C., Botstein, D. & Kahn, R. A. *Proc. natn. Acad. Sci. U.S.A.* **87**, 1238-1242 (1990).
- Dascher, C. & Balch, W. E. *J. Biol. Chem.* **269**, 1437-1448 (1994).
- Zhang, C.-j. et al. *J. Cell Biol.* **124**, 289-300 (1994).
- Kahn, R. A., Kern, F. G., Clark, J., Gelmann, E. P. & Rulka, C. J. *J. Biol. Chem.* **266**, 2606-2614 (1991).
- Donaldson, J. G., Cassel, D., Kahn, R. A. & Klausner, R. D. *Proc. natn. Acad. Sci. U.S.A.* **89**, 6408-6412 (1992).
- Palmer, D. J., Helms, J. B., Beckers, C. J. M., Orci, L. & Rothman, J. E. *J. Biol. Chem.* **268**, 12083-12089 (1993).
- Donaldson, J. G., Finazzi, D. & Klausner, R. D. *Nature* **360**, 350-352 (1992).
- Helms, J. B. & Rothman, J. E. *Nature* **360**, 352-354 (1992).
- Helms, J. B., Palmer, D. J. & Rothman, J. E. *J. Cell Biol.* **121**, 751-760 (1993).
- Orci, L. et al. *Nature* **362**, 648-652 (1993).
- Elazar, Z. et al. *J. Cell Biol.* **124**, 415-424 (1993).
- Glick, B. S. & Rothman, J. E. *Nature* **326**, 309-312 (1987).

43. Pfanner, N. et al. *Cell* **59**, 95–102 (1989).
44. White, J. M. *Science* **258**, 917–924 (1992).
45. Rothman, J. E. & Warren, G. *Curr. Biol.* **4**, 220–233 (1994).
46. Klausner, R. D., Donaldson, J. G. & Lippincott, S. J. *J. Cell Biol.* **116**, 1071–1080 (1992).
47. Mellman, I. & Simons, K. *Cell* **68**, 829–840 (1992).
48. Tanigawa, G., Orci, L., Amherdt, M., Helms, J. B. & Rothman, J. E. *J. Cell Biol.* **123**, 1365–1371 (1993).
49. Stow, J. L. et al. *J. Cell Biol.* **114**, 1113–1124 (1991).
50. Donaldson, J. G., Kahn, R. A., Lippincott, S. J. & Klausner, R. D. *Science* **254**, 1197–1199 (1991).
51. Block, M. R., Glick, B. S., Wilcox, C. A., Wieland, F. T. & Rothman, J. E. *Proc. natn. Acad. Sci. U.S.A.* **85**, 7852–7856 (1988).
52. Tagaya, M., Wilson, D. W., Brunner, M., Arango, N. & Rothman, J. E. *J. biol. Chem.* **268**, 2662–2666 (1993).
53. Whiteheart, S. W., Rossnagel, K., Buhrow, S., Jaenicke, R. & Rothman, J. E. *J. Cell Biol.* **126**, 945–954 (1994).
54. Wilson, D. W. et al. *Nature* **339**, 355–359 (1989).
55. Paquet, M. R., Pfeffer, S. R., Burczák, J. D., Glick, B. S. & Rothman, J. E. *J. biol. Chem.* **261**, 4367–4370 (1986).
56. Diaz, R., Mayorga, L. S., Weidman, P. J., Rothman, J. E. & Stahl, P. D. *Nature* **339**, 398–400 (1989).
57. Beckers, C. J. M., Block, M. R., Glick, B. S., Rothman, J. E. & Balch, W. E. *Nature* **339**, 397–398 (1989).
58. Graham, T. R. & Emr, S. D. *J. Cell Biol.* **114**, 207–218 (1991).
59. Clary, D. O. & Rothman, J. E. *J. biol. Chem.* **265**, 10109–10117 (1990).
60. Clary, D. O., Griff, I. C. & Rothman, J. E. *Cell* **61**, 709–721 (1990).
61. Whiteheart, S. W. et al. *Nature* **362**, 353–355 (1993).
62. Kaiser, C. A. & Schekman, R. *Cell* **61**, 723–733 (1990).
63. Griff, I. C., Schekman, R., Rothman, J. E. & Kaiser, C. A. *J. biol. Chem.* **267**, 12106–12115 (1992).
64. Wilson, D. W., Whiteheart, S. W., Wiedmann, M., Brunner, M. & Rothman, J. E. *J. Cell Biol.* **117**, 531–538 (1992).
65. Söllner, T., Bennett, M. K., Whiteheart, S. W., Scheller, R. H. & Rothman, J. E. *Cell* **75**, 409–418 (1993).
66. Söllner, T. et al. *Nature* **362**, 318–324 (1993).
67. Südhof, T. C. & Jahn, R. *Neuron* **6**, 665–677 (1991).
68. Bennett, M. K. & Scheller, R. H. A. *Rev. Biochem.* **63**, 63–100 (1994).
69. Bennett, M. K., Calakos, N. & Scheller, R. H. *Science* **257**, 255–259 (1992).
70. Inoue, A., Obata, K. & Akagawa, K. *J. biol. Chem.* **267**, 10613–10619 (1992).
71. Oyler, G. A. et al. *J. Cell Biol.* **109**, 3039–3052 (1989).
72. Trimble, W. S., Cowan, D. M. & Scheller, R. H. *Proc. natn. Acad. Sci. U.S.A.* **85**, 4538–4542 (1988).
73. Baumert, M., Maycox, P. R., Navone, F., De Camilli, R. & Jahn, R. *EMBO J.* **8**, 379–384 (1989).
74. Schiavo, G. et al. *Nature* **359**, 832–835 (1992).
75. Hess, D. T., Slater, T. M., Wilson, M. C. & Skene, J. H. P. *J. Neurosci.* **12**, 4634–4641 (1992).
76. Bergeron, J. J. M., Ehrenreich, J. H., Siekevitz, P. & Palade, G. E. *J. Cell Biol.* **59**, 73–88 (1973).
77. Huttner, W. B. *Nature* **365**, 104–105 (1993).
78. Bennett, M. K. & Scheller, R. H. *Proc. natn. Acad. Sci. U.S.A.* **90**, 2559–2563 (1993).
79. Hardwick, K. G. & Pelham, H. R. B. *J. Cell Biol.* **119**, 513–521 (1992).
80. Sogaard, M. et al. *Cell* **78**, 937–948 (1994).
81. Calakos, N., Bennett, M. K., Peterson, K. E. & Scheller, R. H. *Science* **263**, 1146–1149 (1994).
82. Pevsner, J. et al. *Neuron* **13**, 353–361 (1994).
83. Landschulz, W. H., Johnson, P. F. & McKnight, S. L. *Science* **240**, 1759–1764 (1988).
84. Goud, B., Salminen, A., Walworth, N. C. & Novick, P. J. *Cell* **53**, 753–768 (1988).
85. Salminen, A. & Novick, P. J. *Cell* **49**, 527–538 (1987).
86. Segev, N., Mulholland, J. & Botstein, D. *Cell* **52**, 915–924 (1988).
87. Rexach, M. F. & Schekman, R. W. *J. Cell Biol.* **114**, 219–229 (1991).
88. Segev, N. *Science* **252**, 1553–1556 (1991).
89. Bourne, H. R. *Cell* **53**, 669–671 (1988).
90. Chavrier, P., Parton, R. G., Hauri, H. P., Simons, K. & Zerial, M. *Cell* **62**, 317–329 (1990).
91. Plutner, H. et al. *J. Cell Biol.* **115**, 31–43 (1991).
92. Brennwald, P. & Novick, P. *Nature* **362**, 560–563 (1993).
93. Dunn, B., Stearns, T. & Botstein, D. *Nature* **362**, 563–565 (1993).
94. Dascher, C., Ossig, R., Gailwitz, D. & Schmitt, H. D. *Molec. cell. Biol.* **11**, 872–885 (1991).
95. Aalto, M. K., Keranen, S. & Ronne, H. *Cell* **68**, 181–182 (1992).
96. Lippincott-Schwartz, S. et al. *Cell* **60**, 821–836 (1990).
97. Donaldson, J. G., Lippincott, S. J., Bloom, G. S., Kreis, T. E. & Klausner, R. D. *J. Cell Biol.* **111**, 2295–2306 (1990).
98. Orci, L. et al. *Cell* **64**, 1183–1195 (1991).
99. Taylor, T. C., Kahn, R. A. & Melançon, P. *Cell* **70**, 69–79 (1992).
100. Ladinsky, M. S. et al. *J. Cell Biol.* **127**, 29–38 (1994).
101. Brown, H. A. et al. *Cell* **75**, 1137–1144 (1993).
102. Roth, T. F. & Porter, K. R. *J. Cell Biol.* **20**, 313–332 (1964).
103. Pearce, B. M. F. *Proc. natn. Acad. Sci. U.S.A.* **73**, 1255–1259 (1976).
104. Stannos, M. A. & Rothman, J. E. *Cell* **73**, 999–1005 (1993).
105. Traub, L. M., Ostrom, J. A. & Kornfeld, S. *J. Cell Biol.* **123**, 561–573 (1993).
106. Bariowe, C. et al. *Cell* **77**, 895–907 (1994).
107. Nakano, A. & Muramatsu, M. *J. Cell Biol.* **107**, 851–863 (1989).
108. Yoshihisa, T., Barlowe, C. & Schekman, R. *Science* **259**, 1466–1468 (1993).
109. Balch, W. E., Kahn, R. A. & Schwaninger, R. *J. biol. Chem.* **267**, 13053–13061 (1992).
110. Peter, F., Plutner, H., Zhu, H., Kreis, T. E. & Balch, W. E. *J. Cell Biol.* **122**, 865–875 (1993).
111. Balch, W. E., McCaffery, J. M., Plutner, H. & Farquhar, M. G. *Cell* **76**, 841–852 (1994).
112. DeBello, W. M., Betz, H. & Augustine, G. J. *Cell* **74**, 947–950 (1993).
113. Geppert, M. et al. *Cell* (in the press).
114. Brose, N., Petrenko, A. G., Südhof, T. C. & Jahn, R. *Science* **256**, 1021–1025 (1992).
115. Almers, W. et al. *Ann. N.Y. Acad. Sci.* **635**, 318–327 (1991).
116. Monck, M. A. & Fernandez, J. M. *J. Cell Biol.* **119**, 1395–1404 (1992).
117. Neher, E. & Zucker, R. S. *Neuron* **10**, 21–30 (1993).
118. Palade, G. E. in *Subcellular Particles* (ed. Hayashi, T.) 64–83 (Ronald, New York, 1959).
119. Rothman, J. E. & Lenard, J. *Trends biochem. Sci.* **9**, 176–178 (1984).
120. Bennett, M. K. et al. *Cell* **74**, 863–873 (1993).
121. McMahon, H. T. et al. *Nature* **364**, 346–349 (1993).
122. Serafini, T. et al. *Nature* **349**, 215–220 (1991).
123. Duden, R., Griffiths, G., Frank, R., Argos, P. & Kreis, T. E. *Cell* **64**, 649–665 (1991).
124. Stenbeck, G. et al. *EMBO J.* **12**, 2841–2845 (1993).
125. Harrison-Lavole, K. et al. *EMBO J.* **12**, 2847–2853 (1993).
126. Lian, J. P. & Ferro-Novick, S. *Cell* **73**, 735–745 (1993).

ACKNOWLEDGEMENTS. I thank F.-U. Hartl, J. Lenard, L. Orci, T. Söllner, G. Tanigawa, G. Warren and F. Wieland for their helpful comments, J. Olesker and W. Bellamy for preparing the manuscript, and G. Palade and L. Orci for generously providing micrographs. Space limitation has precluded citation of all primary references, which are included here in reviews cited.

Mantle plumes and episodic crustal growth

M. Stein^{*†} & A. W. Hofmann[†]

^{*} Institute of Earth Sciences, The Hebrew University, Givat Ram, Jerusalem, Israel

[†] Max-Planck-Institut für Chemie, Postfach 3060, 55020 Mainz, Germany

Many hitherto puzzling features of the isotope and trace-element geochemistry of the Earth's mantle and crust can be explained if Earth history is punctuated by episodes of enhanced exchange between the lower and upper mantle. Such episodes would replenish the upper mantle with trace elements, and also cause rapid growth of continental crust. This picture is consistent with recent geophysical models in which two-layer convection alternates with episodes of penetrative or whole-mantle convection.

AN understanding of the origin and growth of continental crust stands as one of the major goals in the study of the Earth's evolution. The available geochronology of crust formation shows time periods when large areas of continental crust were formed, alternating with apparently more quiescent periods of low crust-formation rates¹. It may be that these 'quiescent' times merely represent missing continental material, because crust has been destroyed and recycled into the mantle^{2,3}, but actual evidence for this process is indirect at best. Reymer and Schubert^{4,5} have drawn attention to the fact that crust formation rates during some geological periods exceeded by far the rate of

$\leq 1 \text{ km}^3 \text{ yr}^{-1}$ that has prevailed during Phanerozoic time. At the very least, their results show that local crust-formation rates in some regions, such as the Arabian-Nubian Shield or the Superior Province (North America), have been as high as (or higher than) the present-day rate of crust formation averaged over the entire globe. Here we use the working hypothesis that continental-crust formation (hereafter abbreviated as 'crust formation') was at least partly episodic, a point of view that has long been advocated by Moorbath⁶.

There is little doubt that some strong link exists between mantle convection and crust formation, via subduction-driven

arc magmatism and arc accretion, plume-driven magmatic 'underplating' of existing crust, or accretion of plume-generated oceanic plateaus. If crust formation was indeed episodic, then this implies that the convective flow pattern should vary between modes where crust formation rates are high and other modes where these rates are low. Recent simulation studies of mantle convection have indicated the possibility of mantle flow patterns that alternate between single-layer (or 'whole-mantle') and two-layer convection^{7,8}. This may happen because the endothermic phase change at 660 km depth tends to stabilize marginally a two-layer convection regime, which breaks down intermittently, and because accumulated, cold subducted slabs may catastrophically break through the transition zone and sink into the deep mantle. There is also seismological evidence that some slabs penetrate the transition zone at 660 km whereas others do not⁹⁻¹¹. There are, in any case, good reasons to think that mantle convection in general is strongly time dependent¹².

We propose here that periods of strongly penetrative two-layer^{13,14} convection lead to areally extensive crust formation by accretion of plume heads, whereas periods of purely two-layer convection are associated with low rates of crust formation by arc accretion. For the sake of brevity and simplicity, we call this the MOMO (mantle overturn and major orogenies) model.

Our model helps to reinterpret and explain several geochemical observations, the meaning of which has been controversial in the past. They are all consistent with significant but limited input of lower-mantle material into the upper mantle, and some of them require such input. These observations are as follows. (1) Initial isotopic compositions of neodymium and strontium for several areally extensive major 'orogens' such as the Pan-African and the Birimian terranes and the Superior Province are surprisingly uniform and intermediate between primitive mantle and incompatible-element-depleted, 'MORB-type' mantle¹⁵⁻²¹ (the term 'incompatible' refers to elements that have low solid/melt partition coefficients in the mantle and are therefore extracted from the mantle and transferred to the crust: MORB stands for mid-ocean-ridge basalt). (2) The trace-element budget of the continental crust requires extraction of incompatible elements from a mantle reservoir that is larger than the upper mantle but smaller than the whole mantle²²⁻²⁴. (3) The noble-gas budget requires the continued presence of primordial ³He and large amounts of radiogenic ⁴⁰Ar in the lower mantle, but also significant gas transfer to the upper mantle²⁵. (4) The isotopic composition of lead is not consistent with the concentration ratios of thorium to uranium in the upper mantle, if the continental crust and upper mantle evolved as a closed system²⁶. The same holds true for other critical trace-element concentration ratios in the upper mantle and continental crust, because they do not add up to possible primitive mantle ratios^{27,28}.

The model: MOMO events and Wilsonian periods

We assume that the mantle-crust system alternates between two modes of convection and dynamic evolution, one approximating a two-layer convective style with nearly complete isolation of the upper layer, the other being characterized by a strongly penetrative convection mode^{13,14} with significant exchange between lower and upper mantle and crust.

The two states are schematically depicted in Fig. 1. In the first mode, plate tectonics operates in a 'conventional' manner with the Wilson cycle²⁹ of opening and closing of oceans, continental drift and collisions. Continental crust formation rates are low and are dominated by arc processes. The upper mantle is isolated from the lower mantle and is progressively depleted in highly incompatible elements (in the approximate order of Cs, Rb, Ba, Th, U, Pb, Nb, Ta, K, La, Ce, Nd, Sr, Sm³⁰). The sink for these elements is partly the continental crust but also the oceanic crust, which is subducted and temporarily stored at the base of the upper mantle. A Wilsonian period of 500 Myr suffices to process and differentiate nearly all of the upper mantle through generation of basaltic oceanic crust and residual lithosphere. On sub-

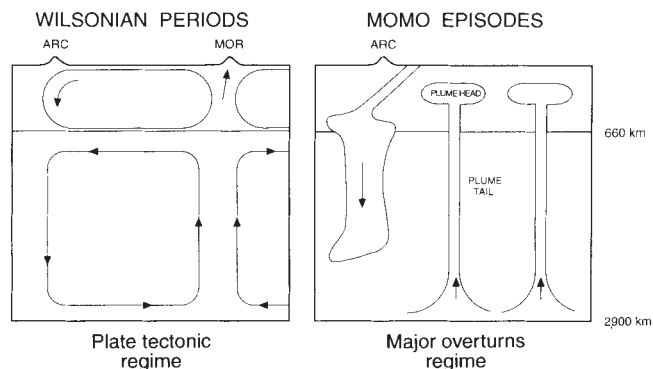


FIG. 1 Schematic representation of the MOMO and Wilsonian²⁹ convection modes. Left, during Wilsonian periods the normal mode of plate tectonics prevails, with opening and closing of oceans, and mantle convection with isolated upper and lower mantle. Plumes originate predominantly from the base of the upper layer, and continental growth is dominated by arc accretion. The upper mantle is progressively depleted in incompatible elements by transfer into the continental crust and temporary storage in subducted oceanic crust. (MOR, mid-ocean ridge.) Right, during MOMO episodes, accumulated cold material descends from the 660 km boundary layer into the lower mantle, and multiple major plumes rise from the core-mantle boundary to the surface, thus creating a major overturn. Large plume heads create oceanic plateaus, some of which may be converted to juvenile continental crust by lateral accretion, or underplate existing continental lithosphere and crust. By entrainment and growth during ascent, the plume heads attain chemical and isotopic compositions similar to that of the bulk residual mantle. This composition is intermediate between that of the severely depleted upper mantle and that of the original, primitive mantle; it is also close to the PREMA composition originally proposed to represent an end-member composition of continental basalts⁵⁵. MOMO episodes cause partial replenishment of incompatible elements in the upper mantle, which had been depleted during the previous Wilson period.

duction, the basaltic crust is assumed to segregate in part at the base of the upper mantle, from where it is recycled to form small plumes^{31,32}. Therefore, the main portion of the upper mantle will become relatively depleted in basaltic component and incompatible elements. Additional depletion is caused by continental growth through extraction of arc magmas in connection with lithospheric subduction.

The other mode (MOMO) involves significant exchange of material between lower and upper mantle (mantle overturn) as well as high rates of continental crust formation (major orogenies), with partial depletion of the lower mantle and partial replenishment of the upper mantle in incompatible elements. We assume this to be effected by strongly penetrative convection with major, possibly 'catastrophic' downwellings and plume upwellings to and from the D' layer at the base of the lower mantle^{7,8,33}. Oceanic plateaus formed by plume heads³⁵ are accreted onto the existing continental crust^{19,20,35} by normal plate-tectonic processes. In addition, high rates of crust formation may be caused by emplacement of continental flood basalts and/or extensive igneous underplating at the base of existing crust (compare ref. 36).

Plumes rising from the lower mantle not only trigger rapid crustal growth, but also replenish the incompatible trace elements in the upper mantle, which had been depleted during the previous Wilsonian period. This happens because some of the plume material does not melt but remains in the upper mantle, and because the initially basaltic oceanic plateaus undergo internal differentiation during accretion to the existing continental crust, so that mafic and ultramafic material is returned to the upper mantle³⁷.

The plume-borne replenishing material needs time to become mixed back into the upper mantle before it manifests itself in the composition of MORB. This is why the most recent event

of major plume intrusions from the lower mantle, which created the large Cretaceous plateaus in the Pacific Ocean (for example, Ontong Java and Manihiki)³⁸, has not noticeably affected the composition of the East Pacific Rise as yet. By the same token, accretion of these plateaus to the continental crust is also delayed until subduction has swallowed the intervening normal oceanic crust and lithosphere. Thus we suggest that the deep-mantle part of a new MOMO event has already taken place, but the Earth's upper-mantle and continental crust 'do not know it yet'.

Our model resembles that of Galer *et al.*²⁴ (in particular their model E1) in some respects. Both models postulate partial convective isolation and partial depletion of the lower mantle by exchange of material with the upper mantle. Because of this, most of the mass balances calculated by these authors apply nearly equally well to our model. The difference lies in the episodic nature of our model, as opposed to the steady-state approach of Galer *et al.*, where entrainment creates a 'leaky' boundary between the two layers.

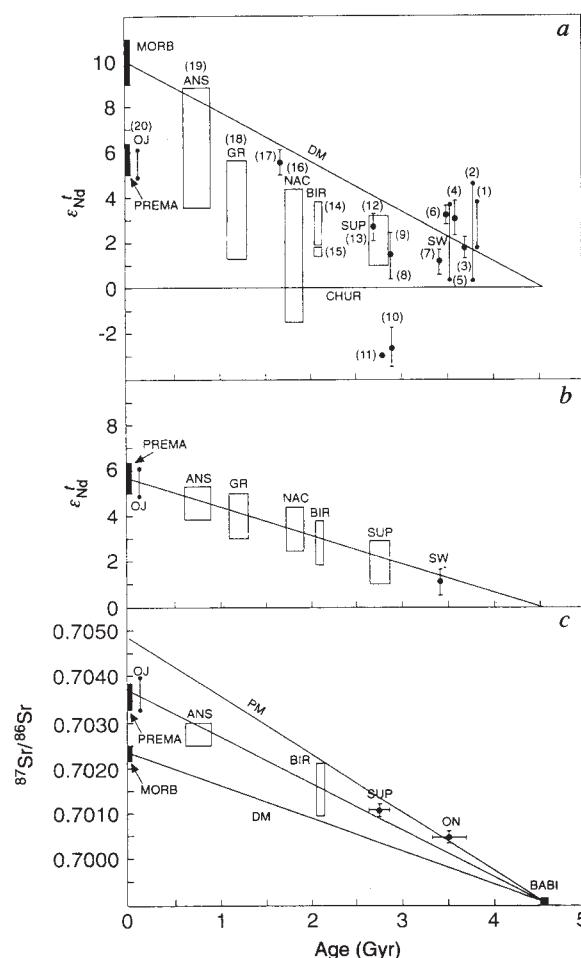
Nd and Sr isotopes

Figure 2a shows a compilation of initial ϵ_{Nd} values for basalts and granitoid rocks as a function of age (ϵ_{Nd} is an expression of the deviation of the $^{143}\text{Nd}/^{144}\text{Nd}$ ratio from the bulk Earth value, in parts per 10⁴). The data on this diagram show considerable scatter, but an approximation to their upper boundary is widely used to describe the evolution of the 'depleted mantle', which is thought to constitute the source of MORB and juvenile continental crust and is therefore used in calculations of crustal model ages³⁹. (The horizontal line represents, by definition, the evolution of the primitive mantle). The scatter seen in this dia-

gram is generally thought to represent variable admixture of older ('recycled') continental material (with $\epsilon_{\text{Nd}} < 0$) to 'juvenile' material derived from the depleted upper mantle (with high ϵ_{Nd}).

The MOMO model implies an alternative source for some of the largest orogens, which correspond to major juvenile crustal provinces. Mean isotopic compositions (and standard deviations) of these are shown in Fig. 2b. Study of the extensive 1.7 to 1.9 Gyr-old Proterozoic crust-forming event of the Northern Hemisphere has shown that the margins of this orogen are contaminated by reworked older crust and thus yield low ϵ_{Nd} values. In its interior, however, this orogen is remarkably uniform in isotopic composition ($\epsilon_{\text{Nd}} = +3$ to $+4$)^{39, 41}. Similarly uniform ϵ_{Nd} values of $+4$ to $+5$ (ref. 16) characterize the northern part of the late Proterozoic (~0.9–0.6 Gyr) Pan-African orogen on the Arabian–Nubian Shield. Only the southern segments (for example, the MORB-type ophiolite of Jabal Gerf) have higher ϵ_{Nd} values between $+5$ and $+8$ (refs 15, 18) and probably represent remnants of normal oceanic crust, most of which was subducted before accretion of the northern Arabian–Nubian Shield. Both basalts and granitoids from the very large Birimian orogen (~2.1 Gyr) of western Africa also have nearly uniform, juvenile isotopic composition ($\epsilon_{\text{Nd}} \approx +3$)^{19, 20}. Other orogens which show similar characteristics are the 1.1–1.4 Gyr Grenville orogen and the 2.7–2.9 Gyr Superior Province from North America^{21, 42–44}. All these initial ϵ_{Nd} values lie on a similar evolution trend (Fig. 2b), which is intermediate between the evolution of the primitive mantle and the more highly depleted upper mantle. It appears that the segments of these orogenies, which show no evidence for significant involvement of older continental crust, have grown from mantle sources that are not as depleted

FIG. 2 Initial ϵ_{Nd} (ϵ_{Nd}^i) and $^{87}\text{Sr}/^{86}\text{Sr}$ isotope ratios as a function of age of Archean, Proterozoic and Phanerozoic magmas from several segments of continental crust. DM, inferred growth line of the depleted upper mantle; CHUR, growth line of chondritic uniform reservoir, PM, primitive mantle reservoir. a, Non-selective compilation of ranges of initial ϵ_{Nd} values from mafic and granitoid rocks of all ages and from all continents. Traditionally, the upper boundary of these scattered data is used to present the isotopic evolution of the upper, depleted mantle, and the lower values are assumed to represent mixtures of this depleted, mantle-derived material and recycled continental crust. Data sources: (1), Akiila gabbros⁵⁶, Greenland; (2), Isua and Amitsoq gneisses^{56, 57}, Greenland; (3), Talga Talga⁵⁸, Australia; (4), Saglek Bay⁵⁹, Labrador; (5), Onverwacht Group⁶⁰, South Africa; (6), Rajasthan⁶¹, India; (7), Swaziland⁶⁰, SW; (8), Yilgarn⁶², Australia; (9), Lewisian⁵⁷, Scotland; (10), Pongola⁶³ and (11), Usushwana⁶³, Swaziland; (12), Kambalda⁶⁴, Australia; (13), Superior Province^{21, 65}, North America, SUP; (14), Birimian^{19, 20}, West Africa, BIR; (15), Guiana⁶⁶, (16), North Atlantic Continent^{39–41}, NAC; (17), Tijeras⁶⁷, USA; (18), Grenville^{42–44, 68}, North America, GR; (19), Arabian–Nubian Shield (Pan-African) (refs 15–18, 69, 70 and M.S. *et al.*, unpublished data), ANS; (20), Ontong Java Plateau³³, Pacific Ocean, OJ; (21), PREMA⁵⁵. b, Initial ϵ_{Nd} in selected major orogens. The data base for each of these orogens is the same as that shown in a, but the data have been filtered to eliminate samples from peripheral regions of the orogens which obviously contain recycled older crust, as in the marginal regions of the Grenville orogen, or ophiolites obviously derived from the upper mantle, as in the Jabal Gerf ($\epsilon_{\text{Nd}} = +8$) in the southern part of the Arabian–Nubian Shield. The sizes of the rectangles indicate the approximate age range and the standard deviation of the ϵ values of each orogen. These orogens form a well defined ϵ_{Nd} -trend, which lies between the growth line of the depleted upper mantle and that of the primitive mantle, and may be interpreted as the approximate growth line of the bulk residual mantle. We interpret each one of the orogenies to be the result of an overturn event in the mantle (MOMO). During such events plume heads rising from the lower mantle mix with, and partly replenish, the upper mantle material in addition to generating major additions to the continental crust. c, Initial $^{87}\text{Sr}/^{86}\text{Sr}$ ratios in several juvenile orogens. Data sources are as defined in a, with the addition of the Onverwacht Group⁷¹, South Africa, ON. The depleted mantle growth curve extends from the value of BABI⁷² ('basaltic achondrite best initial'), thought to be the best estimate of the $^{143}\text{Nd}/^{144}\text{Nd}$ ratio of the initial, undifferentiated Earth to that of average present-day N-MORB²³.



as the upper mantle should be, if it had evolved in a manner implied by the mantle evolution curves found in the literature^{45,46}.

We emphasize that intermediate isotopic signatures, such as those shown in Fig. 2b, are also produced by island arc magmatism as a result of admixture of sediments to the magma source, and in some cases, for example, the Aleutians, the isotopic compositions of the resulting volcanic rocks are remarkably uniform⁴⁷. This has even led to the suggestion that island arc basalts have OIB-like sources⁴⁸ (OIB stands for ocean island basalt). We suggest, however, that this similarity has helped to obscure the fundamental distinction between the two types of igneous processes, and has led further to the almost complete neglect of mantle plumes as significant contributors to the growth of the continental crust (for notable exceptions, see refs^{19,20,35}). Trace-element data, particularly when they include niobium and/or tantalum, could potentially provide more sensitive discriminants between arc-type and plume-type sources of basalts from juvenile crustal provinces. This criterion was successfully used in the case of the Birimian¹⁹, but data are generally of poor quality or lacking for other juvenile crustal provinces.

The input from the lower mantle, implied by the MOMO model, is consistent not only with the growth curve of the major orogens shown in Fig. 2b but also with the evolution line of the depleted mantle (Fig. 2a). DePaolo⁴⁹ noted that the slope of this line is inconsistent with inferred ¹⁴⁷Sm/¹⁴⁴Nd ratios of present-day MORB sources, which would cause ϵ_{Nd} to rise much too steeply with time, and he concluded that the depleted upper mantle must be partially buffered by recycled continental crust. Patchett and Chauvel⁵⁰ noted that this might instead be accomplished by input of primitive mantle material, for example through gradual lowering of the boundary between the upper and lower mantle. Galer *et al.*²⁴ showed that this discrepancy could also be resolved by continuous exchange of material between lower and upper mantle reservoirs. The model proposed here accounts for the highly depleted upper-mantle material by rapid depletion during Wilsonian periods and replenishment during MOMO episodes.

The Sm/Nd and ¹⁴³Nd/¹⁴⁴Nd ratios of the upper mantle are decoupled from one another²² and, in our model, are substantially determined by input from the lower mantle, approximately as described by Galer *et al.*²⁴, but in a highly episodic manner rather than a quasi-steady state. The minimum Sm/Nd ratio of the present-day upper mantle can be estimated from the ratios found in depleted MORB, because the Sm/Nd ratio of a basaltic melt is equal to or less than that of its peridotitic source. ¹⁴⁷Sm/¹⁴⁴Nd ratios as high as 0.23 are common in MORB²². Such high ratios cannot exist in the upper mantle for very long, because (if left in a closed system) they will cause the ϵ_{Nd} to grow twice as fast as the depleted-mantle line shown in Fig. 2a.

The Sr isotopic evolution is shown in Fig. 2c. Again, the mean initial ⁸⁷Sr/⁸⁶Sr ratios of the Pan-African, Birimian and Superior orogens lie on a line which is intermediate between the growth curve of a primitive mantle reservoir and an assumed linear evolution from primordial Sr toward present-day, depleted-mantle (MORB-type) Sr. Addition of lower-mantle material to the depleted upper mantle during MOMO periods will raise Rb/Sr and ⁸⁷Sr/⁸⁶Sr in the upper mantle, but during periods of two-layer convection, ⁸⁷Sr/⁸⁶Sr will remain nearly constant because of the Rb depletion. Because the growth of ⁸⁷Sr/⁸⁶Sr in the upper mantle is irreversible (at least in the absence of an even lower-Rb/Sr reservoir), the present-day ⁸⁷Sr/⁸⁶Sr ratio of the MORB-source reservoir constrains the input of lower-mantle material with higher ⁸⁷Sr/⁸⁶Sr into the upper mantle. Galer *et al.*²⁴ found a total, time-integrated exchange between the two reservoirs amounting to about 30% of the mass of the lower mantle, and this is approximately valid for the present model, because the net result of multiple episodic input is nearly identical to that of steady-state input.

Trace-element budget

Several workers have recognized that the upper depleted mantle cannot be the sole source of the continental crust^{22, 24,26,27} although the trace-element budget of the two reservoirs are complementary in a general way³⁰. Hofmann²³ calculated mass balances between continental crust and a residual mantle consisting of both MORB and OIB source reservoirs on the basis of Nb/U and Ce/Pb ratios, and found that the mass of this residual reservoir is 35–75% of the total mantle. Zindler and Hart²² obtained similar figures ($48 \pm 8\%$ and $65 \pm 17\%$) using Sm/Nd ratios for two different estimates of the composition of the upper mantle. This means that the upper-mantle layer (above 660 km), which amounts to 30% of the total mantle, is insufficient to supply the trace-element budget of the continental crust. Thus neither simple whole-mantle convection nor complete convective isolation between upper and lower mantle could have been maintained throughout the Earth's history. On the other hand, the present model can easily accommodate these mass-balance constraints.

A more difficult problem is presented by the remarkable similarity in Nb/U and Ce/Pb ratios of MORB and OIB, and the contrast to Nb/U and Ce/Pb of the continental crust. Hofmann *et al.*⁵¹ argued that MORB and OIB sources are both residual to the continental crust and that this combined residual reservoir must have been approximately homogenized before its differentiation into depleted (MORB source) and enriched (OIB source) reservoirs. At first sight, this appears to be inconsistent with the difference in composition between upper and lower mantle implied by the MOMO model: the lower mantle, and plumes originating there, should have Nb/U and Ce/Pb ratios different from their upper-mantle values, and closer to primitive-mantle values. However, the issue is not quite this simple. We argue that plumes will ordinarily sample the lower boundary layer of the mantle, which may well consist of recycled lithospheric material (presumably brought down during a previous MOMO episode). If this is true, plumes would be expected to have upper-mantle Nb/U and Ce/Pb ratios. Only the plume heads, which have entrained large amounts of lower-mantle material⁵², may then show the more 'primitive' ratios. The Ontong Java plateau, which probably represents the least contaminated (by continental crust or lithosphere) and best-analysed⁵³ plume head, still yields equivocal information on this point. Nb/U ratios average about 35 ± 13 (1σ) and are thus consistent with being intermediate between an upper-mantle value of 47 and a primitive-mantle value of 30, but they scatter excessively, probably because U is affected by submarine alteration. But the Nb/Th average of 12.3 ± 1.5 (1σ) shows much less scatter and also lies between typical MORB values of about 20 and a primitive-mantle value of about 8, and is therefore consistent with the MOMO model.

Rare-gas budget

The depleted MORB source reservoir has surprisingly high and uniform ratios of primordial ³He to radiogenic ⁴He, and many OIB sources have even higher relative concentrations of ³He. We suggest that this helium is replenished during MOMO episodes and is then progressively depleted. During this depletion process, decay of Th and U continues to produce ⁴He and thus gradually lower the ³He/⁴He in the upper mantle. Unfortunately, this process cannot be constrained by mass-balance calculations because the initial amount of He in the Earth is not known and He is not conserved in the atmosphere. The amount of primordial argon is also not known, but radiogenic ⁴⁰Ar can be calculated from potassium abundance, and argon is not significantly lost from the atmosphere. The mass balance calculated for ⁴⁰Ar shows that about 50% of the total radiogenic argon produced during Earth's history remains in the mantle today²⁵. Most of this must reside in the lower mantle because the upper mantle and continental crust are known to contain only small amounts. Again we would expect the upper mantle to be replenished in

argon during MOMO episodes, and subsequently depleted during Wilsonian periods.

The Th/U versus $^{208}\text{Pb}^*/^{206}\text{Pb}^*$ budget

The ratio of thorogenic to uranogenic lead may be expressed as $^{208}\text{Pb}^*/^{206}\text{Pb}^* \equiv (^{208}\text{Pb}_{\text{total}} - ^{208}\text{Pb}_{\text{primordial}}) / (^{206}\text{Pb}_{\text{total}} - ^{206}\text{Pb}_{\text{primordial}})$. Galer and O'Nions²⁶ pointed out that the Th/U ratio of the MORB source is not consistent with $^{208}\text{Pb}^*/^{206}\text{Pb}^*$ if this reservoir was created by, and therefore has the same mean age as, the formation of the continental crust. This also implies that the upper mantle contains a much larger amount of ('unsupported') thorogenic lead than could have been created in about 2 Gyr (the approximate mean age of the continental crust) by the amount of thorium in the present-day upper mantle. This observation is important because it constitutes hard evidence against the hypothesis that the continental crust and the upper mantle could have evolved as a closed system. Galer and O'Nions²⁶ and Galer *et al.*²⁴ proposed that the upper mantle is actually in an approximate steady-state of input of Th, U and Pb from the lower, less-depleted mantle and output of these elements to the continental crust, and they invoked different residence times for Th, U and Pb to explain the discrepancy between the Th/U ratio inferred from Pb isotopes and the actual Th/U of the upper mantle. They suggested further that the input of lower-mantle material into the upper mantle might occur by entrainment of lower-mantle material into plume sources at the boundary layer between upper and lower mantle (at 660 km depth). As explained above, the MOMO model of intermittent input from the lower mantle yields equivalent results for the mean Th-U-Pb concentrations as does the steady-state model. The low Th/U ratios in present-day MORB sources are simply the result of relative Th depletion during the current Wilsonian cycle.

The PREMA connection

We have postulated⁵⁴ that plume heads may be fossilized beneath, and incorporated into, the continental lithosphere, and that these plume heads represent an average isotopic composition of the original plume source, plus the material entrained from the overlying mantle. Their composition may therefore be close to that of the mean mantle. We suggested that this process may cause the so-called PREMA ('prevalent mantle') composition⁵⁵. Wörner *et al.*⁵⁵ used this term to characterize continental basalts, which converge on relatively 'depleted' Sr and Nd isotopic composition but never reach the degree of depletion that characterizes normal MORB sources. In our view, these basalts are derived from fossilized plume heads that have been trapped at the base of the lithosphere and are tapped by continental rifting. Figure 2 shows that the Nd and Sr growth lines of the major orogens (the MOMO lines) evolve to present-day values of PREMA (as defined in ref. 55). Stein *et al.*¹⁶ demonstrated that the juvenile crust that comprises the Arabian-Nubian Shield is characterized by initial Nd and Sr isotope ratios that are very similar to those of the overlying Phanerozoic alkali basalts. They suggested that the continental crust and the basalts were derived from a PREMA-type plume head that was originally emplaced beneath oceanic lithosphere. The same argument can be made for the West African Birimian Province, where the basalts and granitoid rocks have identical initial isotopic compositions²⁰.

As noted above, the basalts from the Ontong Java oceanic plateau have isotopic compositions very similar to PREMA⁵³, and they may be the best representative of a PREMA-type plume head, which has not been contaminated by continental crust or lithosphere. This should provide opportunities to test the MOMO model. Thus the isotopic comparisons between the Ontong Java basalts and those from the juvenile crustal provinces shown in Fig. 2 should be extended to trace-element patterns. The database for this is quite rudimentary at present: the flat rare-earth-element patterns of the Birimian basalts¹⁹ are indeed similar to those of the Ontong Java basalts⁵³, and the

same holds for the Nb/Ce ratios, which range from 0.3 to 0.5 for Ontong Java basalts and from 0.2 to 1.0 for Birimian basalts. Unfortunately, the Th/U data for the former basalts are too scattered to be useful, and are non-existent for the Birimian basalts. In any case, the MOMO model and its connection to PREMA can be tested in the future by analysing basalts from other oceanic plateaus and other juvenile crustal provinces, not only for isotopic compositions but also, and especially, for the trace elements Nb, Ta, Th, U, Ba and the rare-earth elements.

The MOMO model is consistent with the isotopic record of those orogenies that constitute major additions to the continental crust. It thus appears to be capable of resolving the long-standing controversy regarding the presence or absence of convective isolation of upper- and lower-mantle reservoirs. It provides a plausible mechanism for creating major additions to the continental crust in a geologically short time span. If all these crustal additions had been created by conventional arc accretion processes, one would have to conclude that oceanic plateaus such as the Ontong Java Plateau either did not exist in the more distant geological past, or that all such plateaus were destroyed by subduction in spite of their considerable crustal thickness. □

Received 7 March; accepted 30 September 1994.

1. Taylor, S. R. & McLennan, S. M. *Rev. Geophys.* (submitted).
2. Armstrong, R. L. *Phil. Trans. R. Soc. A* **301**, 443–472 (1981).
3. Gurnis, M. & Davis, G. F. *Geology* **14**, 396–399 (1986).
4. Reymer, A. & Schubert, G. *Geology* **14**, 299–302 (1986).
5. Reymer, A. & Schubert, G. *Tectonics* **3**, 63–77 (1984).
6. Moorbath, S. *Phil. Trans. R. Soc. A* **288**, 401–412 (1978).
7. Machel, P. & Weber, P. *Nature* **350**, 55–57 (1991).
8. Tackley, P. J., Stevenson, D. J., Glatzmaier, G. A. & Schubert, G. *Nature* **361**, 699–704 (1993).
9. Fukao, Y., Obayashi, M., Inoue, H. & Neubai, M. *J. geophys. Res.* **97**, 4809–4822 (1992).
10. van den Hilst, R., Engdahl, R., Spakman, W. & Nolet, G. *Nature* **353**, 37–43 (1991).
11. Zhou, H. W. & Clayton, R. W. *J. geophys. Res.* **95**, 6829–6851 (1990).
12. Christensen, U. *Geophys. Res. Lett.* **14**, 220–223 (1987).
13. Silver, P. G., Carlson, R. W. & Olson, P. *Ann. Rev. Earth planet. Sci.* **16**, 477–541 (1988).
14. Olson, P., Silver, P. G. & Carlson, R. W. *Nature* **344**, 209–215 (1990).
15. Reischmann, T. *Geology* (submitted).
16. Stein, M., Hofmann, A. W. & Goldstein, S. L. *Proc. Gen. Assembly Int. Ass. Volcanol. Chem. Earth's Interior (IAVCE)* (abstr.) 105 (IAVCE, Canberra, Australia, 1993).
17. Stern, R. J. & Kröner, A. *J. Geol.* **101**, 555–574 (1993).
18. Zimmer, M., Kröner, A., Jochum, K. P., Reischmann, T. & Todt, W. *Chem. Geol.* (in the press).
19. Abouchami, W., Boher, M., Michard, A. & Albarède, F. *J. geophys. Res.* **95**, 17605–17629 (1990).
20. Boher, M., Abouchami, W., Michard, A., Albarède, F. & Arndt, N. T. *J. geophys. Res.* **97**, 345–369 (1992).
21. Machado, N., Brooks, C. & Hart, S. R. *Geochim. cosmochim. Acta* **50**, 2335–2348 (1986).
22. Zindler, A. & Hart, S. A. *Rev. Earth planet. Sci.* **14**, 493–571 (1986).
23. Hofmann, A. W. *Phil. Trans. R. Soc. A* **328**, 425–439 (1989).
24. Galer, S. J. G., Goldstein, S. L. & O'Nions, R. K. *Chem. Geol.* **75**, 252–290 (1989).
25. Allègre, C. J., Staudacher, T. & Sarda, P. *Earth planet. Sci. Lett.* **81**, 127–150 (1986).
26. Galer, S. J. G. & O'Nions, R. K. *Nature* **316**, 778–782 (1985).
27. Saunders, A. D., Norry, M. J. & Tarney, J. J. *Petrol.* (spec. lithosphere issue) 415–445 (1988).
28. Sun, S.-S. & McDonough, W. F. in *Magmaism in the Ocean Basins* (eds Saunders, A. D. & Norry, M. J.) 313–345 (Spec. Publ., 42, Geological Soc., Oxford, 1989).
29. Wilson, J. T. *Nature* **211**, 676–681 (1966).
30. Hofmann, A. W. *Earth planet. Sci. Lett.* **90**, 297–314 (1988).
31. Hofmann, A. W. & White, W. M. *Earth planet. Sci. Lett.* **57**, 421–436 (1982).
32. Christensen, U. R. & Hofmann, A. W. *J. geophys. Res.* (in the press).
33. Ringwood, A. R. *Proc. Gen. Assembly Int. Ass. Volcanol. Chem. Earth's Interior (IAVCEI)* (abstr.) 91 (IAVCEI, Canberra, Australia, 1993).
34. Davies, G. F. *Earth planet. Sci. Lett.* **99**, 94–109 (1990).
35. Ben-Avraham, Z., Nur, A., Jones, D. & Cox, A. *Science* **213**, 47–54 (1981).
36. Voshage, H., Hofmann, A. W. & Mazzucchelli, M. *Nature* **347**, 731–736 (1990).
37. Arndt, N. T. & Goldstein, S. L. *Tectonophysics* **163**, 201–212 (1989).
38. Larson, R. L. *Geology* **19**, 547–550 (1991).
39. Patchett, P. J. & Arndt, N. T. *Earth planet. Sci. Lett.* **78**, 329–338 (1986).
40. Patchett, P. J. & Kouvo, O. *Contr. Miner. Petrol.* **92**, 1–12 (1986).
41. Chauvel, C., Arndt, N. T., Kielinzcuk, S. & Thom, A. *Can. J. Earth Sci.* **24**, 396–406 (1987).
42. Marcantonio, F., McNutt, R. H., Dickinson, A. P. & Heaman, L. M. *Chem. Geol.* **83**, 297–314 (1990).
43. Daly, J. S. & McLelland, J. M. *Geology* **19**, 119–122 (1991).
44. McLelland, J. M., Daly, J. S. & Chiarenzelli, J. J. *Geol.* **101**, 97–105 (1993).
45. DePaolo, D. J. *Geochim. cosmochim. Acta* **44**, 1185–1196 (1980).
46. Goldstein, S. L., O'Nions, R. K. & Hamilton, P. J. *Earth planet. Sci. Lett.* **70**, 221–236 (1984).
47. von Drach, V., March, B. D. & Wasserburg, G. J. *Contr. Miner. Petrol.* **92**, 13–34 (1986).
48. Morris, J. D. & Hart, S. R. *Geochim. cosmochim. Acta* **47**, 2015–2033 (1983).
49. DePaolo, D. J. *Geophys. Res. Lett.* **10**, 705–708 (1983).
50. Patchett, P. J. & Chauvel, C. *Geophys. Res. Lett.* **11**, 151–153 (1984).
51. Hofmann, A. W., Jochum, K.-P., Seufert, M. & White, W. M. *Earth planet. Sci. Lett.* **79**, 33–45 (1986).
52. Griffiths, R. W. & Campbell, I. H. *Earth planet. Sci. Lett.* **99**, 66–78 (1990).

53. Mahoney, J. J., Storey, M., Duncan, R. A., Spencer, K. J. & Pringle, M. in *The Mesozoic Pacific: Geology, Tectonics and Volcanism* (eds Pringle, M. S., Sager, W. W., Sliter, W. V. & Stein, S.) 233–261 (Am. Geophys. Union, Washington DC, 1993).
54. Stein, M. & Hofmann, A. W. *Earth planet. Sci. Lett.* **114**, 193–209 (1992).
55. Wörner, G., Zindler, A., Staudigel, H. & Schmincke, H.-U. *Earth planet. Sci. Lett.* **79**, 107–119 (1986).
56. Bennett, V. C., Nutman, A. P. & McCulloch, M. T. *Earth planet. Sci. Lett.* **119**, 299–317 (1993).
57. Hamilton, P. J., O'Nions, R. K., Bridgwater, D. & Nutman, A. *Earth planet. Sci. Lett.* **85**, 105–116 (1983).
58. Gruau, G. et al. *Earth planet. Sci. Lett.* **85**, 105–116 (1987).
59. Stecher, O., Carlson, R. W., Shirey, S. B., Bridgwater, D. & Nielson, T. *Terra Cognita* **6**, (abstr.) 236 (1986).
60. Carlson, R. W., Hunter, D. R. & Barker, F. *Nature* **305**, 701–704 (1983).
61. Macdougall, J. D., Gopalan, K., Lugmair, G. W. & Roy, A. B. in *Workshop on a Cross Section of Archean Crust* (eds Ashwal, L. D. & Card, K. D.) 55–56 (Tech. Rep. 83-03, Lunar and Planetary Inst. Houston, 1983).
62. Fletcher, I. R., Rosman, K. J. R., Williams, I. R., Hickman, A. H. & Baxter, J. L. *Precamb. Res.* **26**, 333–361 (1984).
63. Hegner, E., Kröner, A. & Hofmann, A. W. *Earth planet. Sci. Lett.* **70**, 267–279 (1984).
64. Chauvel, C., Dupre, B. & Jenner, G. A. *Earth planet. Sci. Lett.* **74**, 315–324 (1985).
65. Shirey, S. B. & Hanson, G. N. *Geochim. cosmochim. Acta* **50**, 2631–2651 (1986).
66. Gruau, G., Martin, H., Leveque, B. & Cadevila, R. *Precamb. Res.* **30**, 63–80 (1985).
67. Nelson, B. K. & DePaolo, D. J. *Geol. Soc. Am. Bull.* **96**, 746–754 (1985).
68. Patchett, P. J. & Ruiz, J. J. *Petrol.* **97**, 685–695 (1989).
69. Duyverman, H. J., Harris, N. B. W. & Hawkesworth, C. J. *Earth planet. Sci. Lett.* **59**, 315–326 (1982).
70. Beyth, M., Stern, R. J., Altherr, R. & Kröner, A. *Lithos* **31**, 103–124 (1994).
71. Jahn, B.-M. & Shih, C.-Y. *Geochim. cosmochim. Acta* **38**, 1679–1689 (1974).
72. Papanastassiou, D. A. & Wasserburg, G. J. *Earth planet. Sci. Lett.* **5**, 361–376 (1969).
73. Ito, E., White, W. M. & Göpel, C. *Chem. Geol.* **62**, 157–176 (1987).

ACKNOWLEDGEMENTS. We thank S. Galer, S. Goldstein, J. Patchett and W. White for helping to improve this manuscript by reviews and discussions. We also appreciate discussions with K. Mezger and S. McLennan.

Three-dimensional structure of a hammerhead ribozyme

Heinz W. Pley, Kevin M. Flaherty & David B. McKay*

Beckman Laboratories for Structural Biology, Department of Structural Biology, Sherman Fairchild Bldg, Stanford University School of Medicine, Stanford, California 94305-5400, USA

The hammerhead ribozyme is a small catalytic RNA motif made up of three base-paired stems and a core of highly conserved, non-complementary nucleotides essential for catalysis. The X-ray crystallographic structure of a hammerhead RNA–DNA ribozyme-inhibitor complex at 2.6 Å resolution reveals that the base-paired stems are A-form helices and that the core has two structural domains. The first domain is formed by the sequence 5'-CUGA following stem I and is a sharp turn identical to the uridine turn of transfer RNA, whereas the second is a non-Watson–Crick three-base-pair duplex with a divalent-ion binding site. The phosphodiester backbone of the DNA inhibitor strand is splayed out at the phosphate 5' to the cleavage site. The structure indicates that the ribozyme may destabilize a substrate strand in order to facilitate twisting of the substrate to allow cleavage of the scissile bond.

THE hammerhead ribozyme is one of the few catalytic RNA motifs to have been identified and characterized. It was first recognized as a sequence motif responsible for self-cleavage in satellite RNAs of certain viruses^{1–3}. The consensus sequence required for activity is predicted to have three duplex stems and a conserved 'core' of two non-helical segments (C₃–A₉ and G₁₂–A₁₄ in Fig. 1), plus an unpaired nucleotide at the cleavage site (dC₁₇ in Fig. 1)^{4,5}. Mutagenesis has substantiated this prediction^{5,6}. *In vitro*, short synthetic oligonucleotides with the hammerhead consensus sequence can enzymatically cleave substrates in *trans*^{7–8}. The hammerhead catalyses a transesterification, cutting the 3',5'-phosphodiester bond between nucleotides 17 and 1.1, and forming a cyclic 2',3'-phosphodiester on nucleotide 17 and a free 5'-hydroxyl on nucleotide 1.1 as products^{2,9}. The 2'-hydroxyl of nucleotide 17 is essential for this reaction¹⁰.

The rate of the chemical cleavage step of hammerhead ribozymes is typically ~1 min⁻¹ (refs 7, 8, 11). The activity depends on the presence of divalent cation in roughly millimolar concentration, with a preference for Mg²⁺ or Mn²⁺; it is thought that a divalent ion participates directly in the cleavage reaction^{7,12}. Substantial effort has been invested in identifying specific functional groups in the core of the hammerhead, whose alteration results in decreased activity (for review see ref. 13). These results have not yet converged upon a unique model for the tertiary structure of the ribozyme or a detailed description of its catalytic mechanism.

Although they will not cleave DNA strands, owing to the absence of a 2'-hydroxyl at the cleavage site, hammerheads with the stem-loop configuration (Fig. 1) are active against substrate strands that have a single ribonucleotide at the cleavage site in an otherwise all-DNA strand¹⁰. Thus, a DNA strand that can base-pair with the RNA strand to form stems I and III, with a single unpaired nucleotide at position 17, is an inhibitor that differs from a substrate only by the lack of a 2'-hydroxyl on the sugar at the cleavage site. In this context, we have crystallized an RNA–DNA ribozyme-inhibitor complex and solved its three-dimensional structure.

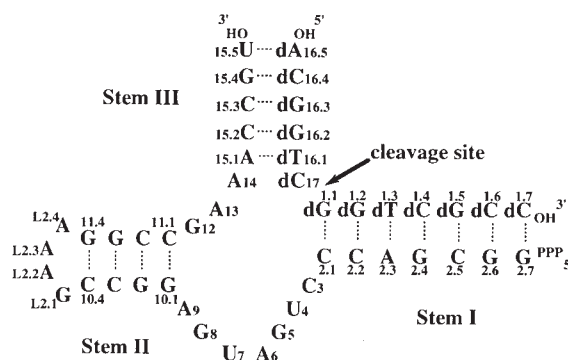


FIG. 1 Sequence, numbering³⁶ and predicted secondary structure of the hammerhead RNA–DNA complex used in this study, with Watson–Crick base pairs shown as dotted lines.

* To whom correspondence should be addressed.

TABLE 1 X-ray structure determination

Spacegroup $P3_221$ $a = 89.7$ Å $c = 185.8$ Å					
Dataset	NAT1	I1	I2	SMCL	NAT2
Detector/ λ	R-axis/1.54 Å	←	Siemens/1.54 Å	→	MAR/1.08 Å
Max. resolution (Å)	2.99	3.09	3.48	3.20	2.60
Reflections/measurements	16,781/29,244	15,433/90,932	9,594/30,908	12,618/48,114	26,816/135,854
Completeness	0.922	0.836	0.700	0.715	0.952
R_{sym}^*	0.027	0.061	0.057	0.054	0.069
MIR analysis (40.0–3.1 Å)					
Number of sites	NA	3	3	3	NA
Phasing power ($=f_{\text{rms}}/E_{\text{rms}}^{\dagger}$)	NA	1.50	1.88	0.98	NA
Figure-of-merit	0.59				
Refinement					
Resolution (Å)	8.0–3.1				8.0–2.6
Reflections	13,958				24,660
Working set	12,547				22,195
Test set	1,411				2,465
R -factor ‡	0.225				0.258
R_{free}	0.286				0.302
Number of atoms	2,994				2,994
R.m.s. bond (Å)	0.012				0.011
R.m.s. angle ($^{\circ}$)	2.99				3.04

Crystallization: Form III crystals were grown from $(\text{NH}_4)_2\text{SO}_4$ as described¹². DNA oligonucleotides, including oligos with 5-I-dU at positions 16.1 or 1.3, were synthesized from commercially available reagents (Glen Research & Milligen) and purified as described¹²; electron spray mass spectrometry was used to identify iodinated versus non-iodinated oligonucleotides in separate peaks from samples chromatographed on monoQ (FPLC; Pharmacia). Data collection and heavy-atom derivatives: Crystals were adapted to a 'mother liquor' consisting of 2.4 M Li_2SO_4 , 10 mM sodium cacodylate (pH 6.0) and 1 mM spermine, and then to a cryosolvent by increasing the percentage glycerol in the mother liquor from zero to 20% in 5% increments at 15-min intervals. Crystals were then flash-frozen in a stream of dry N_2 gas at ~ 100 K. Iodine derivatives I1 and I2 were made by co-crystallization of the RNA strand with DNA strands iodinated at positions 16.1 and 1.3 respectively; the SMCL derivative was made by replacing Mg^{2+} in the mother liquor with 20 mM SmCl_3 ; crystals with Mn^{2+} or Cd^{2+} bound were made by replacing Mg^{2+} in the mother liquor with 100 mM MnCl_2 or CdSO_4 . In the initial search for heavy-atom derivatives, data to 6.2 Å were collected on a Siemens multiwire area detector; reflections were integrated from raw data frames with XDS³⁰. Standard crystallographic computations were carried out with PROTEIN³¹. Heavy-atom sites for both iodine derivatives were solved using the direct methods option of SHELX-86 (ref. 32); the SMCL derivative was solved by difference Fourier. Heavy-atom sites were refined and MIR phases were computed by standard methods³³. The figure-of-merit of phases determined with three derivatives to 6.2 Å resolution was 0.79. Once a set of suitable heavy-atom derivatives were identified, native data (NAT1) were collected to 3.0 Å resolution on a Rigaku R-axis image plate system; integrated intensities were extracted with DENZO (Z. Otwinowski). Derivative data were extended to the diffraction limit of derivative crystals with a Siemens multiwire detector. Subsequently, native data (NAT2) were collected to 2.6 Å on beamline 7-1 of the Stanford Synchrotron Radiation Laboratory on an MAR image plate and processed with DENZO. Model building and refinement: To determine a molecular envelope, a poly(rC) model was built into the RNA strand and a poly(dC) model into the DNA strand of all three copies of the molecule in the asymmetric unit, and an envelope was computed as the boundary of the model with a 3.0 Å radius for each atom; with this envelope, the computed solvent content of the unit cell was 61%. Solvent-flattened phases were computed as before³⁴. The resulting solvent-flattened map was of sufficient quality to trace the entire phosphoribose backbone and to differentiate between purine and pyrimidine bases; position of the iodines in the two iodinated nucleotides confirmed placement of the sequence (Fig. 2a). Refinement using XPLOR³⁵ was initially done with scattering amplitudes from NAT1 to 3.1 Å, and is currently being extended to 2.6 Å with amplitudes from NAT2 collected at SSRL. The present model includes three hammerhead molecules with 47 nucleotides each. The β - and γ -phosphates of the RNA strands were not visible in any of the molecules. No water molecules or metal ions have been placed in the native model. Coordinates will be deposited in the Brookhaven Protein Data Bank. NA, not applicable.

* $R_{\text{sym}} = \sum_{hkl} \sum_{\text{obs}} (I_{\text{obs}} - \langle I \rangle_{hkl}) / \sum_{hkl} \sum_{\text{obs}} \langle I \rangle_{hkl}$, where I_{obs} is the measured diffraction intensity and $\langle I \rangle_{hkl}$ is the mean value of all intensity measurements of (h, k, l) reflection.

$^{\dagger} f_{\text{rms}} = [\sum_{hkl} f_h^2 / n]^{1/2}$; $E_{\text{rms}} = [\sum_{hkl} (F_{\text{PH}} - |F_p + f_h|)^2 / n]^{1/2}$, where f_h is the calculated heavy-atom structure factor; F_p is the positive square root of measured native intensity; F_{PH} is the positive square root of measured derivative intensity; n is the number of reflections contributing to the sum; and $F_p = F_p \exp(i \text{ times calculated native phase})$ and $f_h = f_h \exp(i \text{ times calculated heavy atom phase})$.

$^{\ddagger} R$ -factor $= \sum_{hkl} | \langle F_{\text{obs}} \rangle - \langle F_{\text{calc}} \rangle | / \sum_{hkl} \langle F_{\text{obs}} \rangle$, where $\langle F \rangle$ is mean of all experimental values of F_p and F_{calc} is the structure factor computed from the atomic model.

Structure and activity

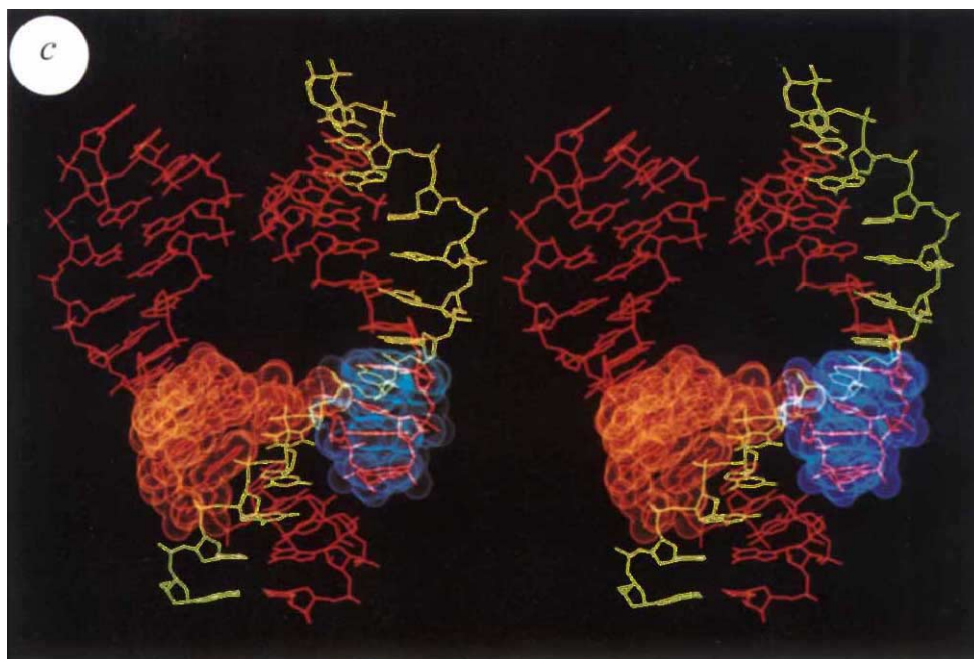
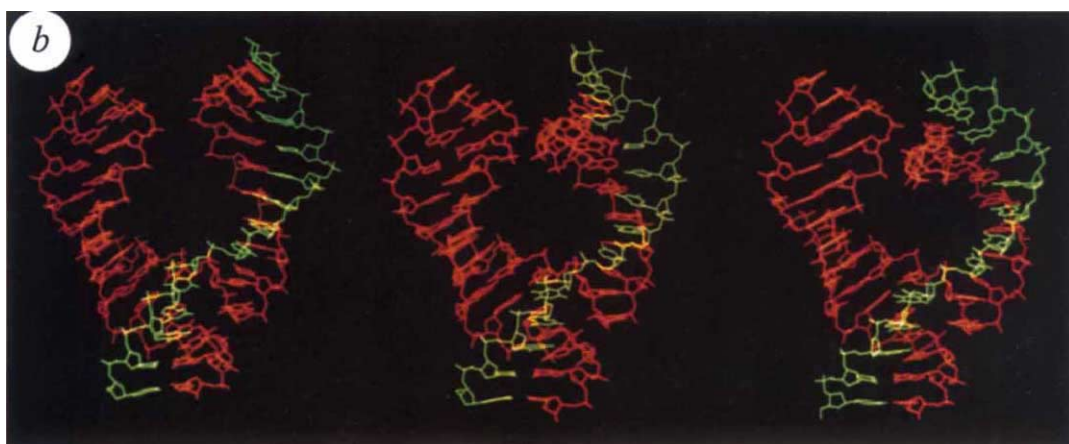
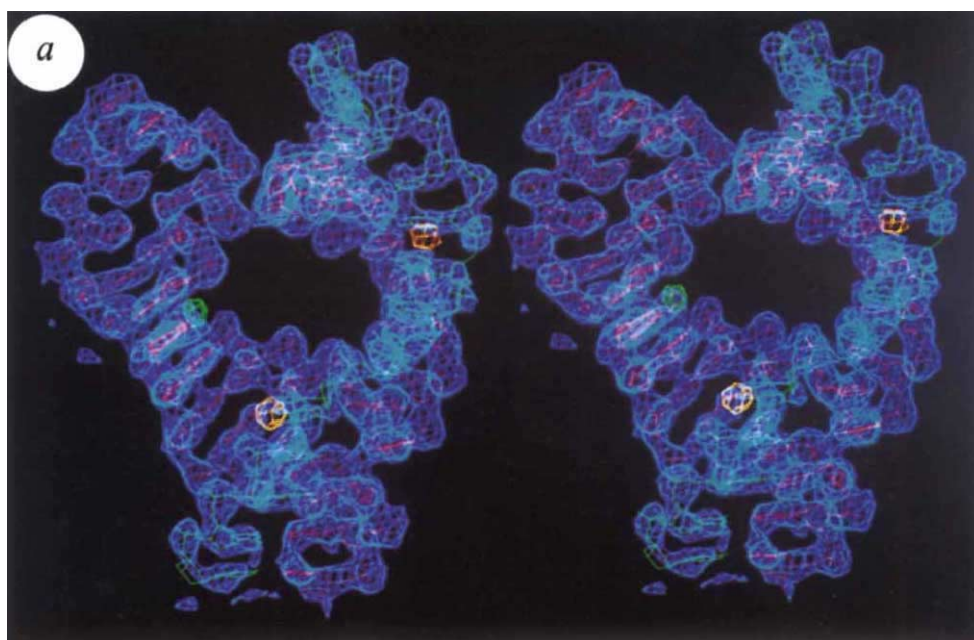
The structure of the hammerhead RNA–DNA complex (Fig. 1) was solved by multiple isomorphous replacement (MIR; Table 1 and Fig 2). Crystals were grown as described with $(\text{NH}_4)_2\text{SO}_4$ as precipitant¹⁴, and then adapted to a stabilizing solution in which 2.4 M Li_2SO_4 replaced $(\text{NH}_4)_2\text{SO}_4$ before data collection, because this substantially increased the resolution limit of the diffraction over that of crystals stabilized in 3.2 M $(\text{NH}_4)_2\text{SO}_4$. To avoid rapid radiation-induced decay at room temperature, crystals were flash-frozen to extend their lifetime indefinitely. There are three copies of the molecule in the crystallographic asymmetric unit; they are in substantially different conformations, so that molecular averaging was not used in the structure determination. Refinement of the model with data to 2.6 Å resolution is in progress.

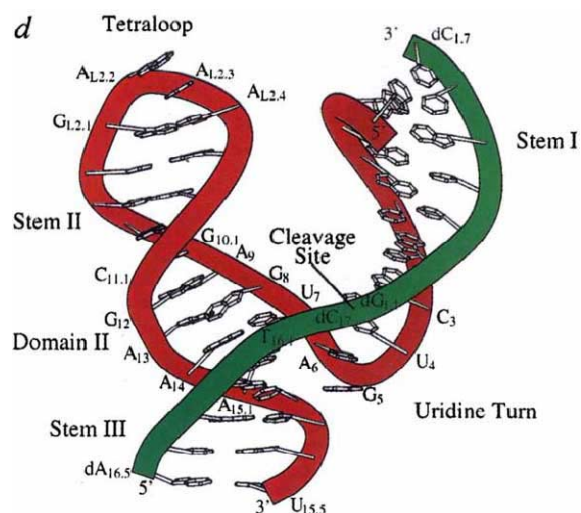
To verify that the ribozyme used in these experiments (Fig. 1) retains activity in high-salt conditions mimicking those of the crystallization, rates of cleavage of (1) the cognate all-RNA substrate strand, and (2) a mostly-DNA substrate with a single ribonucleotide at position 17, were determined under single-turnover conditions (0.5 nM substrate with ≥ 500 nM ribozyme) in the presence of 50 mM Tris-HCl, 10 mM MgCl_2 , and compared with the rates under conditions that were the same apart from inclusion of 1.6 M Li_2SO_4 . For the RNA substrate, the

cleavage rate in the absence of Li_2SO_4 is 2.2 min^{-1} and 0.32 min^{-1} in the presence of 1.6 M Li_2SO_4 . The activity in high salt was dependent on the species of divalent ion present: when MgCl_2 was replaced with 100 mM MnCl_2 in 1.6 M Li_2SO_4 , the cleavage rate of the all-RNA substrate was $>15 \text{ min}^{-1}$. For the mostly-DNA substrate with a single ribonucleotide at the cleavage site, rates in the absence and presence of Li_2SO_4 were 0.008 min^{-1} and 0.0035 min^{-1} respectively. Thus in the presence of 1.6 M Li_2SO_4 , cleavage rates are of the same order as those under standard low-ionic-strength conditions; the hammerhead is active in the high-salt conditions used to determine the structure.

Overall tertiary structure. The hammerhead ribozyme has a tertiary structure not unlike that of a wishbone (Fig. 2). In the view shown in Fig. 2, duplex stems I and II are in close juxtaposition on the upper right and upper left respectively, whereas stem III is at the bottom; all three stems are A-form helices. There is a sharp turn between stems I and II, whereas stems II and III are nearly collinear. The DNA strand is splayed out between nucleotides dT_{16.1} and dC₁₇, so that the first five nucleotides (dA_{16.5}–dT_{16.1}) base-stack on each other and follow stem III, with the remainder of the strand (dC₁₇–dC_{1.7}) base-stacking and following stem I. Although the folding pattern and tertiary interactions are the same in all three crystallographically independent

FIG. 2 Tertiary structure of the hammerhead ribozyme. *a*, Electron density maps. Blue: Fourier map contoured at 1.0σ with a $1.5\text{ \AA}$ cover radius around the atoms of molecule 2 using model-independent, solvent-flattened MIR phases. Orange: difference Fourier maps for derivatives 11 and 12, showing positions of iodine atoms. Green: anomalous difference Fourier map from crystals soaked in 100 mM Mn^{2+} , contoured at 5.7σ . Present model is also shown, with the RNA strand in red and the DNA strand in green. *b*, View of the three hammerhead molecules in the asymmetric unit (numbers 1, 2 and 3, left to right), rotated into similar orientations. The RNA strand is in red, with the 5' end at the upper right and the 3' end at the bottom in each molecule. The DNA strand is in green, with the 5' end at the bottom and the 3' end at the upper right. *c*, Stereo view of molecule 2, with dot surfaces for the non-duplex nucleotides in the core. Dot surface around C_3-A_6 and dC_{17} is blue; around U_7-A_9 and $G_{12}-A_{14}$ it is gold. *d* (facing page), Schematic drawing, with the phosphoribose backbone shown as a ribbon. RNA strand, red; DNA strand, green.



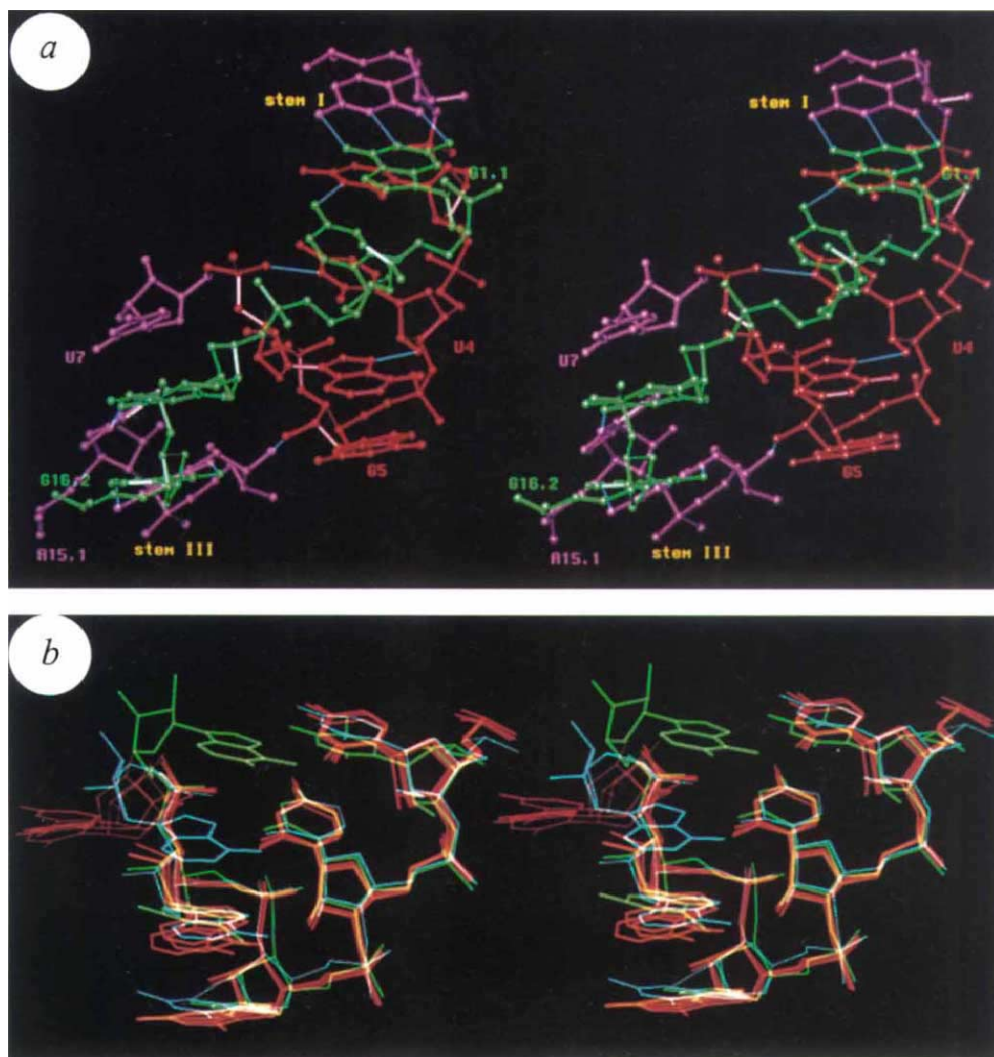


copies of the hammerhead, the molecule has substantial flexibility. This can be seen in Fig. 2*b* as differences in the gap between stems I and II. Quantitatively, the average root-mean-square (r.m.s.) difference in atomic position between pairs of molecules is 3.0 Å. The different conformations of the three molecules will be described in detail elsewhere (H.W.P., K.M.F. and D.B.McK., manuscript in preparation).

The central core of the hammerhead, which is highly conserved in sequence and required in its entirety for catalytic activity, has two structural moieties or domains (Fig. 2*c*). Nucleotides C₃–A₆ (domain 1) make a sharp turn following stem I, whereas nucleotides U₇–A₉ form a non-Watson–Crick duplex with G₁₂–A₁₄ (domain 2).

Domain 1. The turn formed by nucleotides C₃–A₆ is identical in conformation to the uridine turn in the anticodon loop and pseudouridine loop of yeast phenylalanine transfer RNA (tRNA^{Phe}; for review, see ref. 15) (Fig. 3). The hammerhead and tRNA uridine turns superimpose within experimental error: the r.m.s. difference in position of identical atoms is 0.32 Å for the three copies of the hammerhead uridine turn when they are superimposed; the mean r.m.s. difference between the two uridine turns in yeast tRNA^{Phe} is 0.66 Å; the mean r.m.s. difference between the three hammerhead uridine turns and the two turns in tRNA is 0.71 Å.

FIG. 3 Domain 1 and neighbouring nucleotides. *a*, Stereo view of molecular model. Nucleotides C₃–A₆ and the 5' phosphate of U7 are red; other RNA nucleotides are magenta; DNA-strand nucleotides are green; probable hydrogen bonds are blue. *b*, Superposition of the uridine turns of the hammerhead (red) with those in the anticodon (blue) and pseudouridine (green) loop of yeast tRNA^{Phe}.



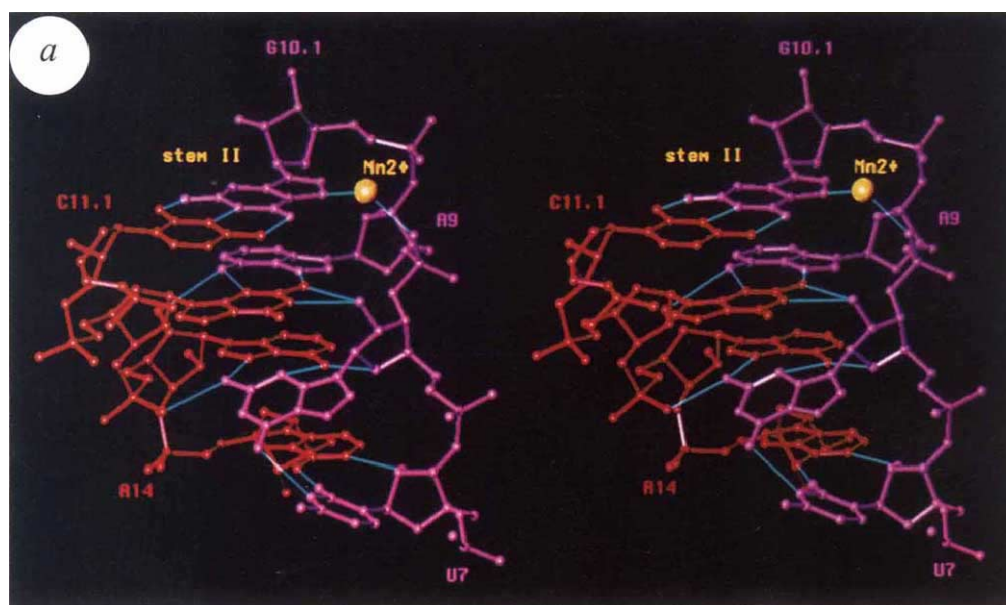
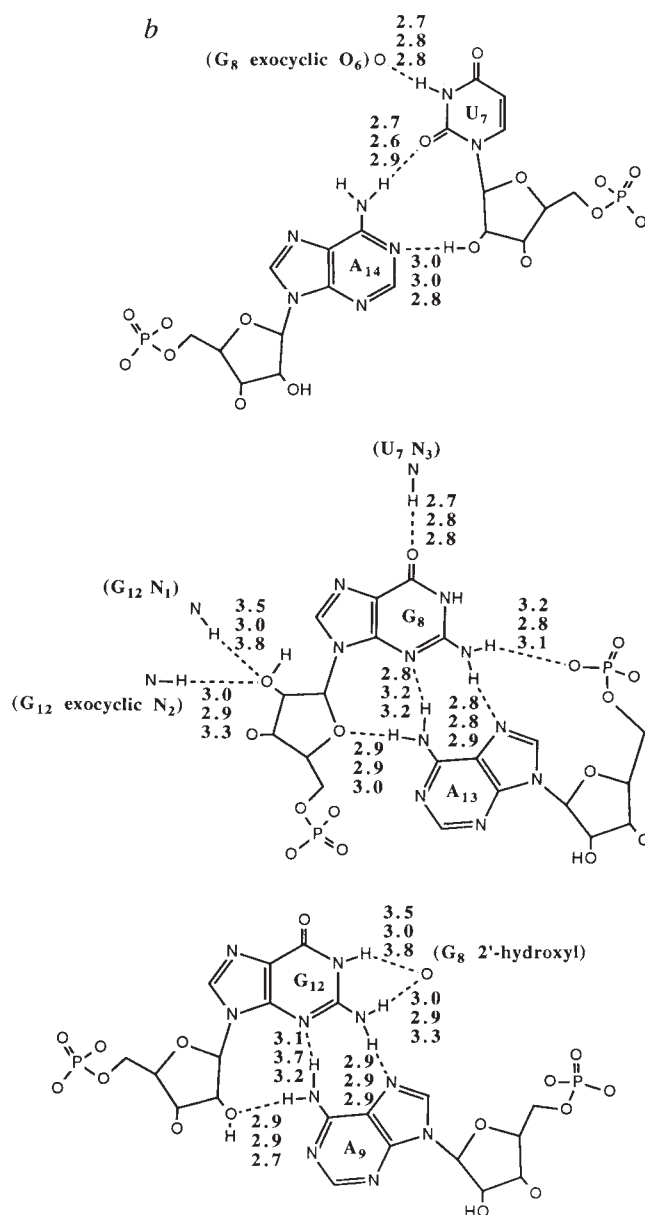


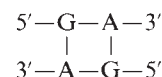
FIG. 4 Domain 2, together with one base pair of stem II. *a*, Stereo view, with U₇–G_{10.1} in magenta and C_{11.1}–A₁₄ in red. Probable hydrogen bonds are blue. Mn²⁺ and its bonds to the *pro*-R_p phosphate oxygen of A₉ and N7 of G_{10.1} are also shown. *b*, Diagrams showing non-Watson–Crick interactions between the nucleotides. The three numbers attached to each bond are bond distances (in Å) between electronegative atoms (nitrogen or oxygen) for molecules 1, 2 and 3 respectively.



Tertiary interactions that stabilize this domain include a hydrogen bond from N3 of U₄ to the *pro*-S_p phosphate oxygen of U₇, as well as a hydrogen bond between the 2'-hydroxyl of U₄ and N7 of A₆. These interactions are consistent with a consensus sequence of UNR for the uridine turn, where N is any nucleotide, R is a purine, and UNR is U₄G₅A₆ in this case. Following the turn, the direction taken by the polynucleotide chain in the hammerhead differs from those after the turns in yeast tRNA^{Phe}. This results from the uridines of the two RNAs hydrogen-bonding two different oxygens on the acceptor phosphate: the hammerhead U₄ hydrogen-bonds the *pro*-S_p phosphate oxygen of U₇ and in yeast tRNA^{Phe}, the hydrogen bonds are to the *pro*-R_p oxygens of the corresponding phosphates.

The 2'-hydroxyl of G₅ is hydrogen-bonded to the 2'-hydroxyl of C_{15.2} in stem III. This hydrogen bond is present in all three copies of the hammerhead, with oxygen–oxygen distances ranging from 2.7 to 2.9 Å; this tertiary interaction of the uridine turn with stem III appears to be important for the hammerhead conformation. Replacing the 2'-hydroxyl of G₅ with other functional groups results in substantial reduction of hammerhead activity^{16–19}; the effect of altering the 2'-hydroxyl of C_{15.2} is not known.

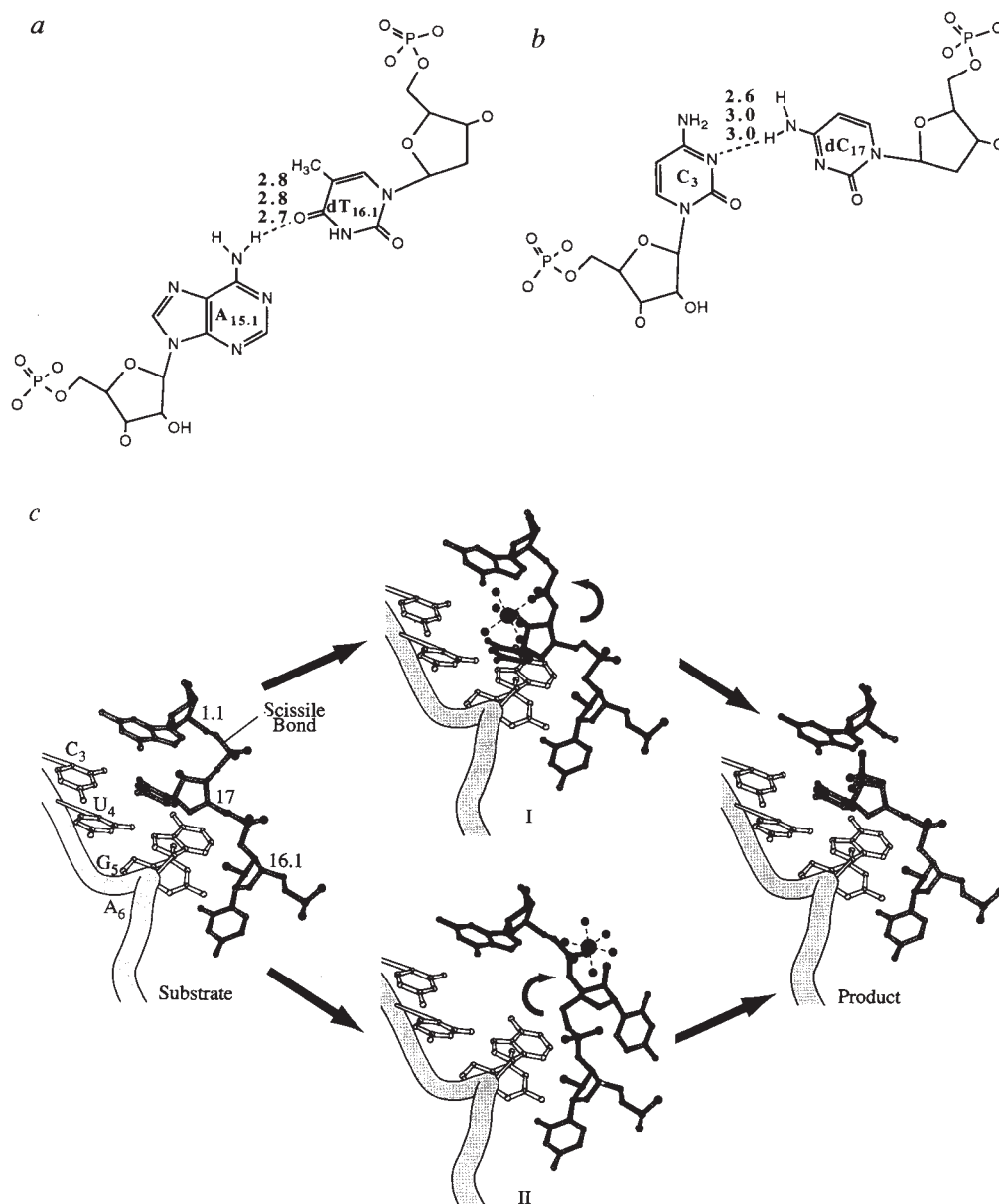
Domain 2. The three nucleotides U₇–A₉ pair with G₁₂–A₁₄ to form a mis-matched duplex (Fig. 4). The base stacking of G₁₂–A₁₄ is roughly parallel and continuous with the base stacking of the stems, with the consequence that there is no major disruption or turn between stems II and III. On the other side of the duplex, the bases of U₇ and G₈ are twisted substantially from parallel stacking. The base pairs G₈–A₁₃ and A₉–G₁₂ form non-Watson–Crick hydrogen bonds (Fig. 4*b*). The base–base hydrogen bond between guanosine and adenine in the G–A mismatches are similar to those observed by NMR²⁰ for a



pair of mismatches in a model duplex of sequence (rGGCGAGCC)₂. The base stacking is also similar, with G₁₂ stacking on G₈, and A₉ stacking above A₁₃. However, the base of G₈ is tilted significantly, and the ribose has a 2'-endo pucker; consequently, the interaction between the 2'-hydroxyl of the G₈ ribose and the non-bonded oxygen of the A₉ phosphate differs substantially from those observed in the model duplex.

Several of the tertiary interactions that we observe in this domain correlate with data on modification of functional groups

FIG. 5 The cleavage site. *a*, Non-Watson-Crick interaction between A15.1 and dT16.1. *b*, Non-Watson-Crick interaction between dC17 and C3. *c*, Hypothetical models of conformational changes that align the ribose 2'-hydroxyl for in-line attack on the phosphorus atom of the scissile bond. The substrate (shaded grey) has been modelled by adding 2'-hydroxyls to the deoxyribose moieties of the nucleotides from the crystallographic structure. The catalytically essential 2'-hydroxyl of nucleotide 17 and the *pro-R_P* non-bonded phosphate oxygen of nucleotide 1.1 are black. An octahedrally coordinated divalent ion, ligating the *pro-R_P* non-bonded phosphate oxygen of nucleotides 1.1, with oxygens representing H₂O molecules at the other coordination positions, is shown in intermediates I and II. A segment of the ribozyme, including bases of the uridine turn, is shown schematically as in Fig. 2d.



that results in decreased hammerhead activity. The 2'-hydroxyl of G₈ hydrogen-bonds to the exocyclic amino group of G₁₂, which is consistent with the substantial decrease in activity caused by altering either of these functional groups^{17-19,21,22}. N3 of U₇ hydrogen-bonds to O6 of G₈; replacing G₈ with 2-aminopurine ribonucleoside¹⁹ (which eliminates O6) or O-6-methylguanosine²³ reduces activity significantly.

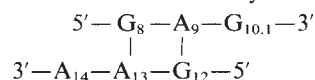
As Mg²⁺ ions are not apparent in the native structure, we soaked both 100 mM Mn²⁺ and 100 mM Cd²⁺ into the crystals and measured the anomalous scattering signal to identify binding sites for divalent ions (Fig. 2a). One strong binding site is observed at essentially the same position in all three copies of the hammerhead. Mn²⁺ and Cd²⁺ both bind at this site by ligating the *pro-R_P* oxygen of the phosphate of A₉ and the N7 atom of G_{10,1} (Fig. 4a). Thiophosphate interference experiments have identified the *pro-R_P* position of the A₉ phosphate as a site whose replacement with sulphur results in decreased Mg²⁺-dependent catalytic activity, suggesting that this atom may interact with divalent ions directly²⁴. Further, replacing G_{10,1} with a pyrimidine, which would eliminate N7 as an available ligand, also results in a substantial decrease in activity^{5,25}. Hence, alteration of either of the functional groups that we observe to ligate Mn²⁺ and Cd²⁺ results in a substantial decrease of Mg²⁺-dependent

activity, suggesting that binding of a divalent ion to this site is important for maximal hammerhead activity. We would not expect Mg²⁺ to bind identically to Mn²⁺ and Cd²⁺ in view of its strong preference for oxygen over nitrogen as a ligand. However, a Mg²⁺ ion could bind near this site with octahedral coordination by directly ligating the *pro-R_P* phosphate oxygens of A₉ and G_{10,1} and by bridging to N7 of G_{10,1} through the hydrogen bond of an H₂O molecule in the first coordination shell of the ion (rather than by ligating the nitrogen directly); we suggest that this may be a Mg²⁺-binding site.

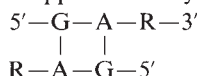
The sequence fingerprint that gives rise to this divalent-ion-binding site,



is a (somewhat distorted) tandem pair of G·A mismatches, followed by a purine, which in this case is G_{10,1}. The resulting conformation of the phosphoribose backbone is such that a divalent ion may ligate one (or possibly two?) phosphate oxygen(s) and the purine N7 simultaneously. With the sequence



the hammerhead has an approximate dyad in this



motif; approximate because opposite sides of the duplex do not have the same conformation. Thiophosphate interference experiments have also identified the *pro*-R_P oxygens of the phosphate of A₁₃ and A₁₄ as atoms whose substitution with sulphur leads to decreased Mg²⁺-dependent activity²⁴ and, by implication, as potential ligands for divalent ions. We do not observe high occupancy of bound metal ions at this site in our crystals, but perhaps under different conditions (for example, lower ionic strength, possibly with some local adjustment of the structure of this domain) this may be an additional binding site for divalent ions.

The cleavage site and catalysis

Several arguments support the proposition that the structure we observe in our crystals is that of an active (and apparently flexible) ribozyme complexed with an inhibitor (the DNA strand). The ribozyme used for structure determination is active in 1.6 M Li₂SO₄, a high-salt condition that mimics the environment in which the crystals are stabilized. It is active against a mostly-DNA substrate with a single ribonucleotide at the cleavage site, albeit with a substantially (~100-fold) reduced turnover rate compared with an all-RNA substrate. Further, as described earlier, many of the tertiary interactions we observe employ functional groups whose modification results in reduced hammerhead activity.

The conformation of the inhibitor in the vicinity of the cleavage site is moulded by its interaction with the ribozyme in general and the uridine turn in particular (Fig. 3a). This inhibitor strand is splayed apart at the phosphodiester linkage 5' to the cleavage site. Each of the two bases on either side of the splayed linkage in the DNA strand makes a single hydrogen bond to its transverse counterpart. A_{15,1}-dT_{16,1} in stem III has a single hydrogen bond between the exocyclic amino group of A_{15,1} and the O4 oxygen of dT_{16,1}, and thus does not form a canonical Watson-Crick base pair (Fig. 5a). The nucleotide at the cleavage site, dC₁₇, forms a single hydrogen bond with the first nucleotide of the core, C₃ (Fig. 5b).

It is well established that cleavage by the hammerhead proceeds with inversion of configuration of the phosphate to form a 2',3'-cyclic phosphate²⁶⁻²⁸; this is probably accomplished by in-line attack of the essential 2'-hydroxyl of the ribose on the phosphodiester bond at the cleavage site. The cleavage requires the presence of divalent ion in millimolar concentration^{7,12,28}. The catalytic rate in the presence of Mg²⁺ is substantially reduced by an R_P thiophosphate at the cleavage site, but not by

an S_P thiophosphate^{27,28}; catalysis with the R_P thiophosphate substrate is restored¹⁰ by Mn²⁺. This suggests that a catalytically essential divalent ion ligates the *pro*-R_P non-bonded oxygen, but not the *pro*-S_P oxygen, of the target phosphodiester bond. It is presumed that the catalytic divalent ion would also interact with other moieties on the ribozyme. It has also been suggested, based on the pH dependence of the catalytic activity, that a hydroxide ion bound to the divalent ion could act as a proton acceptor in the reaction²⁹. Thus, both in-line attack by the ribose 2'-hydroxyl and participation of a specifically bound divalent ion at the active site are thought to be crucial for catalysis.

The conformation we observe at the cleavage site is not one that would allow in-line attack of the 2'-hydroxyl on the phosphodiester bond. Further, we do not see a divalent ion bound at the active site. Thus, if a substrate strand were initially to bind the ribozyme in the same manner as the inhibitor strand we observe in our crystals, a significant conformational rearrangement would be required in the vicinity of the cleavage site both to poise the scissile bond for in-line attack and to form a binding site for the catalytic divalent ion. In principle, this could be accomplished in either of two ways (Fig. 5c). Either the phosphate could twist 'inwards' (intermediate I; Fig. 5c) or the nucleotide could flip 'outwards' (intermediate II; Fig. 5c); each change (or a combination in which changes of both types contribute) could align the 2'-hydroxyl for in-line attack. The former scheme (I), with the phosphate rotating inwards, has the advantage that if a divalent ion were to bind the *pro*-R_P oxygen of the phosphate, it would face the ribozyme. This model identifies several sites in the uridine turn that could either directly ligate a metal ion or hydrogen-bond to H₂O molecules in its first coordination shell at this point. Although our description is only schematic, we note that candidates include the 2'-hydroxyl of U₄, the exocyclic amino group of A₆, and O6 of G₅. In the last scheme (II), in which the nucleotide is flipped out, a divalent ion binding in a similar manner would face the solvent, precluding interaction with the ribozyme.

In the proposed reaction scheme, the function of the catalytic core of the ribozyme would be twofold. First, it would destabilize, or flex, the substrate strand to allow the scissile phosphodiester linkage to twist into a conformation that would allow cleavage. This would presumably be a dynamic process; the apparent flexibility of the ribozyme, and the orientation of stems I and III to kink the inhibitor strand near the cleavage site, suggest that a substrate strand could be twisted considerably in the active site region without major disruption of the stems or the tertiary structure of the core. Second, the core presumably assists in positioning a divalent metal ion so as to facilitate catalysis, but how this occurs remains to be determined. □

Received 23 August; accepted 12 October 1994.

- Prody, G. A., Bakos, J. T., Buzayan, J. M., Schneider, I. R. & Bruening, G. *Science* **231**, 1577-1580 (1986).
- Hutchins, C. J., Rathjen, P. D., Forster, A. C. & Symons, R. H. *Nucleic Acids Res.* **14**, 3627-3640 (1986).
- Buzayan, J. M., Gerlach, W. L. & Bruening, G. *Nature* **323**, 349-353 (1986).
- Forster, A. C. & Symons, R. H. *Cell* **49**, 211-220 (1987).
- Ruffner, D. E., Stormo, G. D. & Uhlenbeck, O. C. *Biochemistry* **29**, 10695-10702 (1990).
- Sheldon, C. C. & Symons, R. H. *Nucleic Acids Res.* **17**, 5679-5685 (1989).
- Uhlenbeck, O. C. *Nature* **328**, 596-600 (1987).
- Haseloff, J. & Gerlach, W. L. *Nature* **334**, 585-591 (1988).
- Buzayan, J. M., Hampel, A. & Bruening, G. *Nucleic Acids Res.* **14**, 9729-9743 (1986).
- Dahm, S. C. & Uhlenbeck, O. C. *Biochimie* **72**, 819-823 (1990).
- Hertel, K. J., Herschlag, D. & Uhlenbeck, O. C. *Biochemistry* **33**, 3374-3385 (1994).
- Dahm, S. C. & Uhlenbeck, O. C. *Biochemistry* **30**, 9464-9469 (1991).
- Bratty, J., Chartrand, P., Ferbeyre, G. & Cedergren, R. *Biochim. biophys. Acta* **1216**, 345-359 (1993).
- Pley, H. W., Lindes, D. S., DeLuca, F. C. & McKay, D. B. *J. biol. Chem.* **268**, 19656-19658 (1993).
- Quigley, G. J. & Rich, A. *Science* **194**, 796-806 (1976).
- Perreault, J. P., Wu, T. F., Cousineau, B., Ogilvie, K. K. & Cedergren, R. *Nature* **344**, 565-567 (1990).
- Fu, D. J. & McLaughlin, L. W. *Proc. natn. Acad. Sci. U.S.A.* **89**, 3985-3989 (1992).
- Williams, D. M., Pieken, W. A. & Eckstein, F. *Proc. natn. Acad. Sci. U.S.A.* **89**, 918-921 (1992).
- Tuschl, T., Ng, M. M., Pieken, W., Benseler, F. & Eckstein, F. *Biochemistry* **32**, 11658-11668 (1993).
- SantaLucia, J. J. & Turner, D. H. *Biochemistry* **32**, 12612-12623 (1993).

- Slim, G. & Gait, M. J. *Biochem. biophys. Res. Commun.* **183**, 605-609 (1992).
- Fu, D. J., Rajur, S. B. & McLaughlin, L. W. *Biochemistry* **32**, 10629-10637 (1993).
- Grasby, J. A., Jonathan, P., Butler, G. & Gait, M. J. *Nucleic Acids Res.* **21**, 4444-4450 (1993).
- Ruffner, D. E. & Uhlenbeck, O. C. *Nucleic Acids Res.* **18**, 6025-6029 (1990).
- Tuschl, T. & Eckstein, F. *Proc. natn. Acad. Sci. U.S.A.* **90**, 6991-6994 (1993).
- van Tol, H., Buzayan, J. M., Feldstein, P. A., Eckstein, F. & Bruening, B. *Nucleic Acids Res.* **18**, 1971-1975 (1990).
- Slim, G. & Gait, M. J. *Nucleic Acids Res.* **19**, 1183-1188 (1991).
- Koizumi, M. & Ohtsuka, E. *Biochemistry* **30**, 5145-5150 (1991).
- Dahm, S. C., Derrick, W. B. & Uhlenbeck, O. C. *Biochemistry* **32**, 13040-13045 (1993).
- Kabsch, W. *J. appl. Crystallogr.* **21**, 916-924 (1988).
- Steigemann, W. thesis, Techn. Univ. München (1974).
- Sheldrick, G. M. *Acta crystallogr.* **A46**, 467-473 (1990).
- Rossmann, M. G. *Acta crystallogr.* **A32**, 774-777 (1976).
- Brandhuber, B. J., Boone, T., Kenney, W. C. & McKay, D. B. *J. biol. Chem.* **262**, 12306-12308 (1987).
- Brunger, A. T., Kuriyan, J. & Karplus, M. *Science* **235**, 458-460 (1987).
- Hertel, K. J. *et al. Nucleic Acids Res.* **20**, 3252 (1992).

ACKNOWLEDGEMENTS. We thank O. Uhlenbeck, D. Herschlag, A. Pardi and N. Usman for discussion, S. Fischer of Hewlett-Packard for mass spectrometry of iodinated oligonucleotides, and members of D. Herschlag's laboratory for help with enzyme assays. This work was supported by a grant to D.B.M. from the NIH, and by the resources of the Beckman Laboratories for Structural Biology. Dataset NAT2 was collected at SSRL which is operated by the Department of Energy, Office of Basic Energy Sciences, Division of Chemical Sciences. The SSRL Biotechnology Program is supported by the NIH Biomedical Resource Technology Program, Division of Research Resources. Further support is provided by the Department of Energy, Office of Health and Environmental Research.

The large-scale distribution of X-ray clusters of galaxies

A. K. Romer^{*†}, C. A. Collins^{*‡}, H. Böhringer[§],
R. G. Cruddace^{||}, H. Ebeling^{§¶}, H. T. MacGillivray[#]
& W. Voges[§]

^{*} Physics Department, University of Durham, South Road,
Durham DH1 3LE, UK

[†] Department of Physics and Astronomy, Northwestern University,
Evanston, Illinois 60208, USA

[‡] Astrophysics Group, Liverpool John Moores University, Byrom Street,
Liverpool L3 3AF, UK

[§] Max-Planck-Institut für Extraterrestrische Physik,
D-8046 Garching bei München, Germany

^{||} Naval Research Laboratory, Washington DC 20375, USA

[¶] Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK

[#] Royal Observatory Edinburgh, Blackford Hill,
Edinburgh EH9 3HJ, UK

CLUSTERS of galaxies are not distributed randomly in space, but are themselves clustered, reflecting inhomogeneities in the early Universe¹. The degree of clustering—usually expressed as a correlation length, which measures the characteristic scale for clustering—can therefore be used to determine the size of the initial density fluctuations that gave rise to the clusters^{1–3}. Optical studies of galaxy clusters^{4–7} have indicated a correlation length that conflicts with the predictions of some theories of large-scale structure formation^{1–3}, leading to the suggestion that these optical samples are biased in that foreground or background galaxies not physically associated with a cluster are counted as part of it^{8–10}. Here we report a measurement of the correlation length for a sample of clusters that were selected based on their X-ray emission, which is free from the bias that is inherent to optical studies. We find a correlation length of 13–15 h^{-1} Mpc, where h is the Hubble constant in units of 100 km s⁻¹ Mpc⁻¹. There is no evidence for clusters being significantly elongated along the line of sight, contrary to previous suggestions^{11,12}.

One of the most controversial results in cosmology over the last decade has been the clustering amplitude of galaxy clusters, as measured by the spatial correlation function (ξ_{cc}). Most estimates of this function have been based on an analysis of the Abell optical-cluster catalogue^{13,14}. These studies indicated that the correlation length of Abell clusters $r_0 \geq 20 h^{-1}$ Mpc and $\xi_{cc} = (r/r_0)^{-\gamma}$, where $\gamma \approx 2$ and r is the spatial separation of cluster pairs^{4–7}. This implies that clusters are at least 12 times more clustered than galaxies on the same spatial scale. However, measurements of the clustering amplitude from the Abell catalogue has proved controversial. There is evidence that this catalogue, which was constructed by visual scans of photographic plates, is biased by several selection effects¹⁵. In addition, measurements of ξ_{cc} from optical cluster samples constructed from digitized galaxy catalogues suggest that $r_0 \approx 15 h^{-1}$ Mpc (refs 16, 17). Despite the improved quality of these digitized cluster samples, there is still no agreement as to why the values of r_0 derived from them should be smaller than those derived from Abell cluster samples^{8–10,18}. The fundamental problem with all these optical studies is that the initial cluster selection relies entirely on identifying projected galaxy enhancements in two dimensions. Therefore, both the Abell and the digitized optical cluster samples suffer from projection effects, whereby clusters become contaminated by foreground and background galaxies. This effect is especially severe when two cluster halos overlap along the line of sight. This results in an overestimate of the number of galaxy members within each cluster (the richness class). Therefore, in any optically selected cluster sample with a minimum richness threshold, there will be a tendency to include clusters with intrinsic richness below this threshold, but which lie in directions towards the halos of rich systems. This in turn

increases the number of close pairs of clusters along the line of sight and so artificially increases the clustering amplitude^{9,10}; it also causes the clustering pattern to become strongly anisotropic in the line-of-sight or redshift-space direction. This anisotropy is clearly present in the poorer (richness $R \geq 0$, see legend to Fig. 2) Abell cluster samples^{8,16} (see inset to Fig. 3) and in the digitized surveys^{16,17}, although its strength may depend on richness⁶. Because of these difficulties, uncertainty remains in the amplitude of ξ_{cc} for optical samples.

In contrast, the cluster selection for our study is based on X-ray emission. This emission originates from hot intergalactic gas ($kT \approx 5$ keV) and is brightest over a region $\sim 250 h^{-1}$ kpc wide in the cluster core. This region is significantly smaller than that over which the optical cluster radius is defined ($>1 h^{-1}$ Mpc). As a result, X-ray cluster samples will be relatively free from the projection effects that may plague optical catalogues. The Rosat all-sky survey¹⁹ presents a perfect opportunity to provide a fresh observational input to the study of X-ray cluster correlations^{20,21}. This is the first uniform all-sky imaging survey performed at X-ray energies (0.1–2.4 keV) and has detected large numbers of X-ray-luminous clusters.

Drawing on X-ray information from the Rosat standard analysis software system (SASS)²² and on digitized optical information from the COSMOS plate-measuring machine²³, we have identified 199 cluster candidates above a SASS X-ray flux limit of 10^{-12} erg cm⁻² s⁻¹ in a 3,100 degree² region of the Rosat all-sky survey ($22 h \leq RA \leq 3 h$, $-50^\circ \leq \text{dec.} \leq 2^\circ$, $|b| \geq -40^\circ$: RA, right ascension; dec., declination; b , galactic latitude). An X-ray source was identified as a potential cluster if there was a significant enhancement, over background level, in the galaxy density in any of five circular cells (radii 1–10 arcmin) centred on its SASS position. In the interests of completeness, any enhancement with a $>80\%$ chance of being real was included in the candidate list. Optical spectra have been taken in order to obtain redshifts for the clusters and to confirm identifications. A total of 38 contaminating objects, such as stars and active galactic nuclei, have been excluded, either by examination of the optical images or on the basis of their spectra. Of the remaining 161 candidates, 80% now have redshifts, 78 from our own observations and 50 from the literature. A more detailed description of the redshift survey and identification procedure will be presented

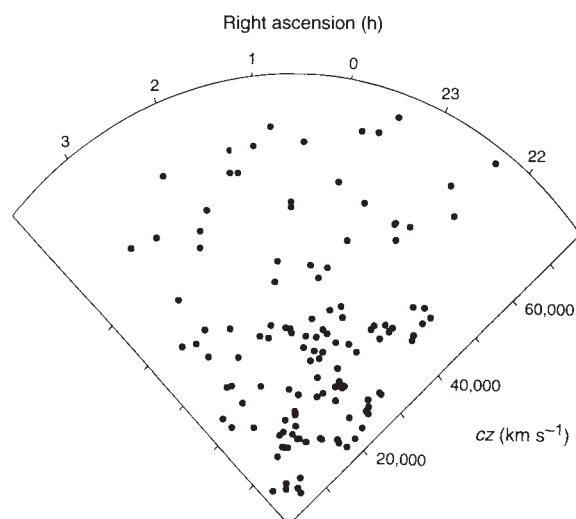


FIG. 1 The right ascension cone diagram for 128 X-ray clusters forming our flux-limited sample over the full declination range $-50^\circ \leq \text{dec.} \leq 2^\circ$ (c is the velocity of light, z is redshift). Our survey contains clusters out to a redshift 60,000 km s⁻¹ ($z \leq 0.2$) and covers a volume $\sim 10^7 h^{-3}$ Mpc³. The cluster distribution shows evidence of voids and superclusters on scales ~ 10 – $50 h^{-1}$ Mpc and is qualitatively consistent with our knowledge of galaxy clustering²⁵.

elsewhere. A right ascension cone diagram for the 128 clusters is shown in Fig. 1.

The correlation function for this sample of clusters was determined by comparing its clustering properties with those of a simulated data set with a random spatial distribution of points. The comparison sample was constructed by choosing coordinates at random from the survey area, and then assigning to each position a redshift drawn from the observed X-ray cluster redshift distribution (with $3,500 \text{ km s}^{-1}$ gaussian smoothing). The amplitude of the correlation function at any point can be estimated using the relation

$$\xi_{cc}(r) = 2 \frac{n_r}{n_c} \frac{N_{cc}}{N_{cr}} - 1 \quad (1)$$

where $N_{cc}(r)$ and $N_{cr}(r)$ are the numbers of the cluster-cluster and cluster-random pairs with spatial separations between r and $r + \Delta r$, and n_c and $n_r (=100n_c)$ are the total numbers of clusters and random points respectively. We used logarithmic bins, to improve the signal-to-noise ratio at large r , with $\Delta \log(r) = 0.2$. For each bin ξ_{cc} was determined 50 times, using different random catalogues, and an average taken. These average values are represented by the filled symbols in Fig. 2. The error bars (1σ) plotted in Fig. 2 were derived using bootstrap statistics²⁴. The bootstrap scheme consists of calculating ξ_{cc} for randomly chosen subsamples of the data (in this case 100 subsamples). This properly takes into account the errors introduced by sampling a finite size of a clustered distribution. A least-squares fit to the points gives

$$\xi_{cc}(r) = \left(\frac{r}{r_0}\right)^{-\gamma}, \quad r_0 = 12.9 \pm 2.2 h^{-1} \text{ Mpc}, \quad \gamma = 1.8 \pm 0.4 \quad (2)$$

We have investigated possible systematic effects in our data by carrying out two tests. (1) We have calculated a least-squares fit to $\xi_{cc}(r)$ for $r \leq 40 h^{-1} \text{ Mpc}$, thus excluding all data points that are consistent with zero clustering ($\xi_{cc}(r) = 0 \pm \delta \xi_{cc}(r)$). This gives $r_0 = 15.1 \pm 2.4 h^{-1} \text{ Mpc}$ and $\gamma = 1.4 \pm 0.4$. (2) To test for contamination of our sample by X-ray-emitting active galactic nuclei, we have calculated $\xi_{cc}(r)$ for the 92 clusters in our sample which show significant angular extent in their X-ray

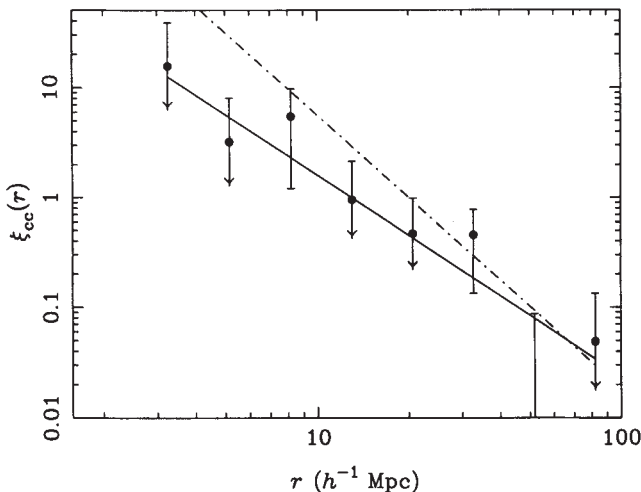


FIG. 2 The spatial correlation function for the 128 clusters in our redshift sample. The solid line shows the best fit to all points (●) out to $85 h^{-1} \text{ Mpc}$, $\xi_{cc}(r) = (r/r_0)^{-\gamma}$, where $r_0 = 12.9 h^{-1} \text{ Mpc}$ and $\gamma = 1.8$. The point at $52 h^{-1} \text{ Mpc}$, $\xi_{cc} = -0.04 \pm 0.07$, is only shown as an upper limit. The dot-dashed line indicates the result of Postman *et al.*⁷, $r_0 = 20 h^{-1} \text{ Mpc}$, $\gamma = 2.4$, for Abell clusters with richness class $R \geq 0$. This richness limit corresponds to all clusters found by Abell with more than 29 galaxies between m_3 and $(m_3 + 2)$, where m_3 is the magnitude of the third brightest galaxy in the cluster. The 1σ error bars shown are derived from bootstrap statistics²⁴.

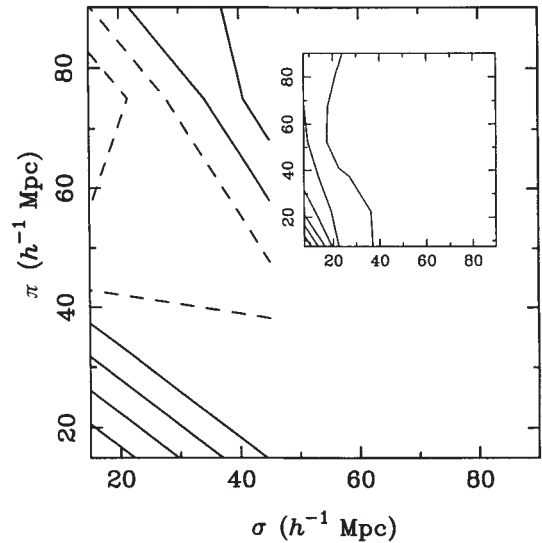


FIG. 3 The contour plot of $\xi_{cc}(\sigma, \pi)$ in the line-of-sight direction π , and in the perpendicular direction σ , for: main figure, the 128 Rosat clusters (contour intervals are 0.11, dashed contours denote negative values); inset, the 351 Abell clusters taken from Postman *et al.*⁷ (contour intervals are 0.84). Note the large elongations in the redshift direction in the inset compared to the main figure.

emission. Fitting over the range $3 - 85 h^{-1} \text{ Mpc}$ we find $r_0 = 13.2 \pm 2.2 h^{-1} \text{ Mpc}$ and $\gamma = 2.2 \pm 0.4$. These results represent the full range of r_0 and γ that we found, and demonstrates the high degree of stability in the X-ray correlation function.

Also shown in Fig. 2 is the correlation function measured by Postman *et al.*⁷ for 351 Abell clusters, $r_0 = 20 \pm 5 h^{-1} \text{ Mpc}$, $\gamma = 2.5 \pm 0.2$. It is clear from this Figure that our correlation function amplitude is systematically smaller than the value measured for Abell clusters. It should be stressed that it is difficult to make a meaningful comparison between our sample, which is X-ray-flux limited, and the optical samples which are volume and galaxy-richness limited. In addition, some of the disparity between the measures of r_0 may be attributable to a reported richness dependence in the correlation function¹⁸. However, the important difference between the X-ray and Abell samples is reflected in Fig. 3. The main part of this Figure shows the contours of the correlation function calculated in two dimensions $\xi_{cc}(\sigma, \pi)$, where $\pi = |cz_1 - cz_2|$ is the separation in the line-of-sight direction; and $\sigma = (r^2 - \pi^2)^{1/2}$ is the separation in the perpendicular direction; the inset shows the identical plot for the Postman sample. The isotropy seen in the main part of Fig. 3, contrasts sharply with the elongations in the redshift direction of the correlation contours seen in the inset.

The absence of a preferred direction of clustering in the X-ray selected sample provides a direct demonstration that the elongations seen in the Abell cluster samples are probably due to projection effects. The only other explanation for this discrepancy would be that Abell clusters and X-ray clusters have intrinsically different clustering properties. This is unlikely, not least because 109 of 128 clusters in our X-ray sample are also Abell clusters. Suggestions made in the past to explain the anisotropy seen in Fig. 3 inset seem much less viable in light of these new results. Our data are consistent neither with superclusters being geometrically elongated along the line-of-sight¹¹ in real space, nor with a "finger-of-God" effect. This is the name given to elongations in redshift space caused by a significant, $\sim 2,000 \text{ km s}^{-1}$, peculiar velocity component to the redshifts in addition to the dominant Hubble expansion¹².

In this Letter we have reported the new observational approach of using an X-ray selected cluster sample to study

large-scale structure, a strategy which is independent of the controversy surrounding the optical results. In the wider context of cosmological implications, our smaller value of r_0 ($15 h^{-1}$ Mpc) removes most of the discrepancy between the correlation amplitude of rich clusters and predictions from the standard model ($\Omega=1$) Universe, filled with cold dark matter^{1,3}. More work is required to determine which models best fit the correlation function beyond this scale. $\square$

Received 5 July; accepted 29 September 1994.

- White, S. D. M., Frenk, C. S., Davis, M. & Efstathiou, G. *Astrophys. J.* **313**, 505–516 (1987).
- Bahcall, N. A. & Cen, R. *Astrophys. J.* **398**, L81–L84 (1992).
- Holtzman, J. A. & Primack, J. R. *Astrophys. J.* **405**, 428–436 (1993).
- Bahcall, N. A. & Soneira, R. M. *Astrophys. J.* **270**, 20–38 (1983).
- Huchra, J. P., Henry, J. P., Postman, M. & Geller, M. J. *Astrophys. J.* **365**, 66–85 (1990).
- Peacock, J. A. & West, M. J. *Mon. Not. R. astr. Soc.* **259**, 494–504 (1992).
- Postman, M., Geller, M. J. & Huchra, J. P. *Astrophys. J.* **384**, 404–422 (1992).
- Efstathiou, G., Dalton, G. B., Sutherland, W. J. & Maddox, S. J. *Mon. Not. R. astr. Soc.* **257**, 125–134 (1992).
- Sutherland, W. *Mon. Not. R. astr. Soc.* **234**, 159–172 (1988).
- Sutherland, W. & Efstathiou, G. *Mon. Not. R. astr. Soc.* **248**, 159–167 (1991).
- Jing, Y.-P., Pilonis, M. & Valdarnini, R. *Astrophys. J.* **389**, 499–509 (1992).
- Bahcall, N. A., Soneira, R. M. & Burgett, W. S. *Astrophys. J.* **311**, 15–24 (1986).
- Abell, G. O. *Astrophys. J. Suppl. Ser.* **3**, 211–288 (1958).
- Abell, G. O., Corwin, H. G. & Olowin, R. P. *Astrophys. J. Suppl. Ser.* **70**, 1–138 (1989).
- Lucey, J. R. *Mon. Not. R. astr. Soc.* **204**, 33–43 (1983).
- Nichol, R. C., Collins, C. A., Guzzo, L. & Lumsden, S. L. *Mon. Not. R. astr. Soc.* **255**, 21p–24p (1992).
- Dalton, G. B., Efstathiou, G., Maddox, S. J. & Sutherland, W. *Astrophys. J.* **390**, L1–L4 (1992).
- Bahcall, N. A. & West, M. J. *Astrophys. J.* **392**, 419–423 (1992).
- Voges, W. in *Proc. Satellite Symp. (Int. Space Year Conf.)* (eds Guyenne, T. D. & Hunt, J. J.) 9–19 (European Space Agency, Paris, 1992).
- Lahav, O., Edge, A. C., Fabian, A. C. & Putney, A. *Mon. Not. R. astr. Soc.* **238**, 881–895 (1989).
- Nichol, R. C., Briel, U. G. & Henry, J. P. *Mon. Not. R. astr. Soc.* **267**, 771–778 (1994).
- Voges, W. et al. in *Proc. Satellite Symp. 3 (Int. Space Year Conf.)* (eds Guyenne, T. D. & Hunt, J. J.) 223–224 (European Space Agency, Paris, 1992).
- Yentis, D. J. et al. in *Digitised Optical Sky Surveys* (eds MacGillivray, H. T. & Thomson, E. B.) 67–75 (Kluwer, Dordrecht, 1992).
- Ling, E. N., Frenk, C. S. & Barrow, J. D. *Mon. Not. R. astr. Soc.* **223**, 21p–27p (1986).
- de Lapparent, V., Geller, M. J. & Huchra, J. P. *Astrophys. J.* **302**, L1–L6 (1986).

ACKNOWLEDGEMENTS. We thank R. C. Nichol, L. Guzzo, G. B. Dalton, H. Andernach and J. P. Huchra for making available unpublished redshifts. The assistance of the staff at the Anglo-Australian Telescope and SAAO Observatory is also warmly acknowledged. We thank R. Nichol for his advice during the reduction and analysis of our data, and R. Mann and A. Heavens for their guidance in the interpretation of these results. C.C. holds a PPARC advanced research fellowship.

Discovery of a nearby spiral galaxy behind the Milky Way

R. C. Kraan-Korteweg*, A. J. Loan†, W. B. Burton‡, O. Lahav†, H. C. Ferguson§, P. A. Henning|| & D. Lynden-Bell†

* Kapteyn Astronomical Institute, Postbus 800, 9700 AV Groningen, The Netherlands

† Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK
‡ Leiden Observatory, Postbus 9513, NL-2300 RA Leiden, The Netherlands

§ Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, Maryland 21218, USA

|| Department of Physics and Astronomy, University of New Mexico, 800 Yale Boulevard NE, Albuquerque, New Mexico 87131, USA

THE disk of the Milky Way contains a lot of gas and dust, which obscures about 20% of the extragalactic sky. Galaxies hidden behind the Milky Way may have an important influence on the dynamics of the Local Group and its peculiar motion relative to the cosmic microwave background radiation^{1,2}. Here we report the discovery of a large spiral galaxy, which we call Dwingeloo 1, during the course of a search for emission from atomic hydrogen (H I) associated with galaxies hidden by the disk of the Milky Way—such H I emission is not obscured by the disk if the velocity of the emission differs from that of the local gas³. The new galaxy seems to be associated with the group containing IC342 and the

Maffei galaxies, and a subsequent optical image suggests that it is of type SBb. The detection of Dwingeloo 1 early in the course of this survey suggests that many more galaxies hidden behind the Milky Way remain to be discovered.

The Dwingeloo obscured galaxy survey (DOGS) is a long-term project to search systematically for galaxies in the northern Milky Way, by detecting H I emission. At present, the 25-m Dwingeloo radio-telescope (built in 1956) is fully dedicated to this project. The receiver at a bandwidth of 20 MHz gives coverage over the velocity range $0 < V_{\text{LSR}} < 4,000 \text{ km s}^{-1}$ (in the frame of the local standard of rest, LSR). Lower velocities have already been surveyed by the Dwingeloo/Leiden Galactic H I survey⁴. The beamwidth of the telescope is 0.6° . The survey points are ordered in a honeycomb pattern with a grid spacing of 0.4° , resulting in nearly 15,000 pointings for the covered galactic latitude strip $|b| \leq 5^\circ$. Our strategy was to first conduct a fast search with 5-minute integrations to uncover the most massive galaxies, followed by a deep survey with 1 hour of integration per pointing to produce a deeper, flux-limited catalogue.

We have started to inspect the scans from the fast blind search in the northern intersection of the supergalactic plane with the galactic plane. There the expectation for detections is large because various extragalactic filamentary structures at different velocities disappear into the obscuration layer of the Milky Way and some very nearby galaxies are already known. Our first inspection of the data revealed an intriguing signal in the spectra

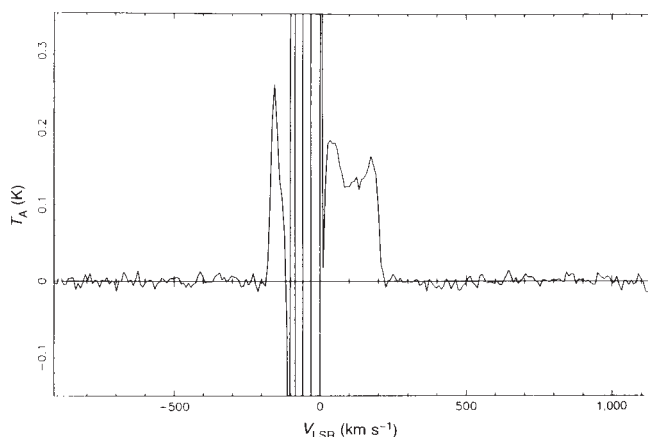


FIG. 1 Spectrum of H I emission from Dwingeloo 1 (Dw1) observed with the Dwingeloo 25-m radio-telescope (55-min integration). The radio-telescope has a 1,024-channel digital autocorrelator spectrometer and a system temperature of $\sim 40 \text{ K}$. The velocity is relative to the local standard of rest, the intensity in terms of the antenna temperature, T_A (1 K $\equiv 7.5 \text{ Jy}$). The instrumental profile (baseline) was minimized by position switching (on-off). The spectrum is displayed after one Hanning smoothing, and residual baseline subtraction. Dw1 has peak intensities at 1.4 and 1.2 Jy, and a full velocity-width (20%) of 200 km s^{-1} . Dw1 displays the classical 'double-horned' profile of an inclined spiral galaxy. Galactic emission disturbs this profile in the range $-170 < V_{\text{LSR}} < +40 \text{ km s}^{-1}$ (measured in both the 'on' and 'off' spectra, resulting in both positive and negative residuals in the 'on-off' spectrum).

of four neighbouring pointings of the radio-telescope, suggesting a galaxy with a large angular extent located within these pointings (at $l \approx 138.5^\circ$, $b \approx -0.1^\circ$). This position coincides with a very low surface brightness feature on the red Palomar sky survey plate. It has an estimated diameter of 2.2 arcmin in a catalogue (G. K. T. Hau, H.C.F., O.L. and D.L.-B., manuscript in preparation) of about 2,500 optically identified galaxy candidates.

After retuning the receiver to place the lower velocities more centrally in the band, renewed observations with the Dwingeloo telescope with a 55-minute integration pointed at the optical counterpart (right ascension (1950.0) 02 h 53 min 01 s, declina-

tion (1950.0) + 58° 42' 38") clearly confirmed the signal. The resulting H I spectrum is shown in Fig. 1. The strong irregular (positive and negative) signal over the range $-170 < V_{\text{LSR}} < +40 \text{ km s}^{-1}$ is dominated by the Milky Way gas. The 'double-horned' feature ($\sim 200 \text{ km s}^{-1}$ wide) centred at $V_{\text{LSR}} \sim +110 \text{ km s}^{-1}$, with peaks at about 1.4 and 1.2 Jy, is the emission from our new galaxy. The signal is clearly not Galactic in origin. Galactic H I does not reach high positive velocities in this part of the Milky Way, and the 'double-horned' profile is a typical signature of spiral galaxies. Dwingeloo 1 (Dw1) lies only 2° away from the spiral galaxy Maffei 2 discovered in the infrared over 25 years ago⁵. Maffei 2 is also seen in our survey but a large fraction of the profile is lost in the Galactic H I. The low-velocity horn of Dw1 overlaps slightly with Galactic H I and the observed linewidth of Dw1 could be a lower limit too. But the symmetry of the signal indicates that most of the emission has been measured (as confirmed below).

To obtain further proof of Dw1 and to estimate its physical and morphological properties, we initiated follow-up observations at various wavelengths. A 12-h observation with the Westerbork synthesis radio-telescope (WSRT) clearly confirms the nearby large galaxy detected with DOGS. From the integrated H I map in Fig. 2, an angular extent of ~ 14 arcmin can be deduced. A distinct deficiency of H I is evident in the centre. This is often seen in barred spiral galaxies. The inclination, as estimated from the integrated H I image, is $i \approx 50^\circ$. The central velocity and rotation velocity agree well with the values deduced from the Dwingeloo spectrum. They are also consistent with the optical spectrum obtained at the William Herschel telescope (La Palma). The WSRT data imply a steep rotation curve in the inner part (± 2 arcmin) and a fairly flat continuation to the outermost H I contours. Further higher-sensitivity observations are being made.

Figure 3 is composite (V, R and I band) image of Dw1 based on observations at the Isaac Newton telescope (La Palma). The galaxy has a distinct bar, from which spiral arms unfold at nearly right angles. The arms can be traced over nearly 180° . The large diameter visible on this image is ~ 4.2 arcmin. The morphology suggested in this image is that of a SBb galaxy. However, the foreground obscuration could still be hiding the outer parts of Dw1, and Dw1 might well be a SBc galaxy comparable to M83.

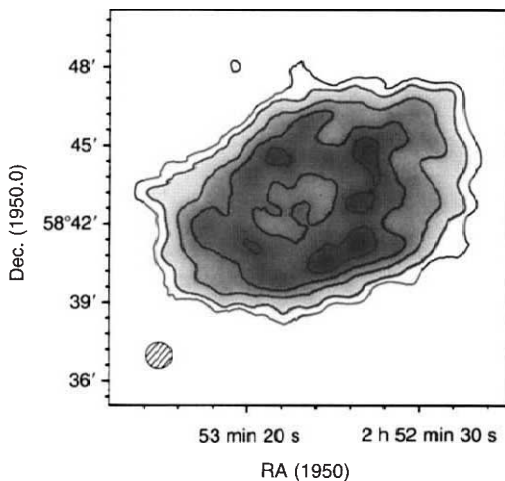


FIG. 2 Integrated H I map of Dw1 (prepared by M. Verheijen, Groningen), obtained with the Westerbork synthesis radio-telescope (12-h integration). The contours display H I column densities of 1.9, 3.2, 6.3, 9.5 and $13 \times 10^{20} \text{ N(H I) cm}^{-2}$. The outermost contour represents a 2σ measurement. The hatched circle marks the size of the FWHM beam. The angular extent of the H I-gas of Dw1 is 14 arcmin. The deficiency of H I emission in the centre is often observed in barred galaxies, as is the ring-like structure surrounding the bar. The inclination $i \approx 50^\circ$.

We now evaluate the distance and the physical parameters of Dw1. The recession velocity derived from the H I spectrum and the WSRT data is $V_{\text{LSR}} = +112 \text{ km s}^{-1}$. Correcting for the motion within our Milky Way, the radial velocity relative to our Galaxy is $V_G = V_{\text{LSR}} + 220 \cos(b) \sin(l) = 258 \text{ km s}^{-1}$. The peculiar velocities of nearby galaxies can be a considerable fraction of their recession velocities. V_G is therefore not indicative of the distance to Dw1 and a redshift-independent distance is required. We use the Tully-Fisher relation between the extinction-corrected (blue) absolute magnitude and the inclination-corrected linewidth of spiral galaxies, $M_B^{o,i} = -6.69 \log \Delta v_{20\%} - 2.77$, which is based on 13 local calibrators⁶. With a full velocity width $\Delta v_{20\%} = 201 \text{ km s}^{-1}$ (Dwingeloo and WSRT data), and an inclination of $i = 50^\circ$ (WSRT map), an absolute total blue magnitude of $M_B^{o,i} = -19.0$ mag, corresponding to a luminosity of $L_B = 6.2 \times 10^9 L_\odot$ (where $L_\odot$ is the solar luminosity), is predicted. The blue magnitude, corrected for galactic and internal extinction (A_B^o and A_B^i , respectively), is $m_B^{o,i} = m_B - A_B^o - A_B^i \approx 8.3$ mag. The blue magnitude is an estimate from the Palomar Sky Survey plate (by R.C.K.-K.), the Galactic extinction is derived from the H I column density^{4,7}. The resulting distance modulus ($m_B^{o,i} - M_B^{o,i}$) ≈ 27.3 mag translates to a distance $D \approx 3$ Mpc. It must be stressed that the large scatter in the blue Tully-Fisher relation ($\sigma = 0.7$ mag) precludes a precise distance determination. Moreover, the magnitude estimate is uncertain and the Galactic extinction should be regarded as a lower limit, because of the unknown extinction from molecular gas. Thus the distance of 3 Mpc should be regarded as a first approximation only. We soon will be able to derive a more accurate distance to Dw1 using the infrared Tully-Fisher relation.

This distance allows us to evaluate the H I mass M_{HI} and the total mass M_{tot} (including dark matter) of Dw1. The intensity of Dw1 is $I = 209 \text{ Jy km s}^{-1}$. This implies⁸ a mass of H I given by

$$M_{\text{HI}} \approx 4.5 \times 10^8 \left(\frac{D}{3 \text{ Mpc}} \right)^2 M_\odot \quad (1)$$

where $M_\odot$ is the solar mass. With the rotation velocity $v_{\text{rot}} = \frac{1}{2} \Delta v / \sin(i) \approx 130 \text{ km s}^{-1}$ we can derive the total mass out to a certain radius r_* . We choose r_* to be the radius where the rotation curve is flat and out to which we still detect the H I emission. This is at an angular radius $\theta \approx 7$ arcmin which corresponds to $r_* \approx 6 \text{ kpc}$ for $D = 3 \text{ Mpc}$. Assuming newtonian dynamics and spherical symmetry,

$$M_{\text{tot}}(< r_*) = \frac{v_{\text{rot}}^2 r_*}{G} \approx 2.5 \times 10^{10} \left(\frac{v_{\text{rot}}}{130 \text{ km s}^{-1}} \right)^2 \left(\frac{D}{3 \text{ Mpc}} \right) \left(\frac{\theta}{7'} \right) M_\odot \quad (2)$$

For comparison, the mass of our Galaxy out to the same radius is about $(220/130)^2 \approx 3$ times larger. The resulting ratios $M_{\text{HI}}/L_B \approx 0.07 (M_\odot/L_\odot)$ and $M_{\text{HI}}/M_{\text{tot}} \approx 0.02$ lie well within the range determined for Sb galaxies.

With the presently available distance and mass estimates we can now address the question of the effect of Dw1 on the very local Universe. Table 1 lists galactic coordinates, extinction in blue, Galactocentric velocity and estimated distance for Dw1, M31, Maffei 1, Maffei 2 and IC342. The spatial proximity of Dw1 to the Maffei galaxies and IC342 supports the idea that it is a member of that group⁹, and not of the Local Group. The boundaries of the Local Group and its membership are ill-defined, but it is believed to be dominated by the Milky Way and Andromeda (M31). Moreover, the motion and direction of Dw1 on the sky imply that Dw1 is not kinematically bound to the Local Group.

TABLE 1 Parameters of Dw1 and neighbouring galaxies

Galaxy	l	b	A_B^0	V_G (km s $^{-1}$)	D (Mpc)	Ref.
M31	121.2°	-21.6°	0.3	-123	0.67	20
Maffei 1	135.8°	-0.6°	6.6	+145	2.1	21
					4.2	22
					3.5	23
Maffei 2	136.5°	-0.3°	5.9	+136	5 ?	24
IC342	138.2°	10.6°	3.1	+178	1.8	25
Dw1	138.5°	-0.1°	5.8	+258	3 ?	This work

Symbols used: $l(b)$, galactic longitude (latitude); galactic extinction; A_B^0 , V_G , radial velocity relative to our Galaxy; D , distance; Ref., source of D data.

The internal dynamics of the Maffei/IC342 group are poorly known. It probably contains about 15 further galaxies, mainly dwarfs¹⁰. The addition of a new massive spiral galaxy to this group is noteworthy. In fact, several authors^{11,12} have speculated, for a variety of reasons, about the existence of a neighbouring galaxy to Maffei 1 and Maffei 2. The total mass of the Maffei/IC342 group including Dw1 is probably larger than the mass of the Milky Way. At a distance of only 3 Mpc, this might well affect the validity of the 'timing argument' which is often applied to yield the mass and age of the Local Group¹³⁻¹⁶. New galaxies will also affect the reconstruction of galaxy orbits back in time^{11,17}, in particular the motions of Maffei 1, Maffei 2 and IC342. In addition, Dw1 should be included when estimating the acceleration vector of the Local Group due to all other galaxies^{18,19} (or the mass distribution they represent). Very nearby galaxies hidden behind the Milky Way (for example IC342, Centaurus A and Dw1) affect both the direction and

convergence of the acceleration vector¹, which again has consequences for the density parameter of the universe Ω_0 .

Dw1 is not likely to be the only galaxy behind the zone of avoidance to have escaped detection. An inventory of more spiral galaxies based on this blind H I search will be available, helping to resolve many of the problems briefly discussed here. □

Received 26 August; accepted 29 October 1994.

1. Kraan-Korteweg, R. C. in *Proc. 2nd DAEC Meeting on the Distribution of Matter in the Universe* (eds Mamon, G. & Gerbal, D.) 202-211 (Observatoire de Paris, Paris, 1993).
2. Lahav, O. in *4th DAEC Meeting* (eds Balkowski, C. & Kraan-Korteweg, R. C. (Astr. Soc. Pacific Conf. Ser., San Francisco, in the press).
3. Kerr, F. J. & Henning, P. A. *Astrophys. J.* **320**, L99-L103 (1987).
4. Burton, W. B. & Hartmann, D. in *Proc. 4th DAEC Meeting* (eds Balkowski, C. & Kraan-Korteweg, R. C.) (Astr. Soc. Pacific Conf. Ser., San Francisco, in the press).
5. Maffei, P. *Publ. astr. Soc. Pacif.* **80**, 618-621 (1968).
6. Kraan-Korteweg, R. C., Cameron, L. M. & Tammann, G. A. *Astrophys. J.* **331**, 620-640 (1988).
7. Burstein, D. & Heiles, C. *Astr. J.* **87**, 1165-1189 (1982).
8. Mihalas, D. & Binney, J. *Galactic Astronomy: Structure and Dynamics* 2nd edn 490 (Freeman, New York, 1981).
9. Van den Bergh, S. *Nature* **231**, 35-36 (1971).
10. Schmidt, K. H. & Boller, T. *Astr. Nachr.* **313**, 189-231 (1992).
11. Peebles, P. J. E. *Astrophys. J.* **362**, 1-13 (1990).
12. Hurt, R. L., Merrill, K. M., Gatley, I. & Turner, J. L. *Astr. J.* **105**, 121-127 (1993).
13. Kahn, F. D. & Woltjer, L. *Astrophys. J.* **130**, 705-717 (1959).
14. Lynden-Bell, D. *Observatory* **101**, 111-114 (1982).
15. Valtonen, M. J., Byrd, G. G., McCall, M. L. & Innanen, K. A. *Astr. J.* **105**, 886-893 (1993).
16. Raychaudhury, S. & Lynden-Bell, D. *Mon. Not. R. astr. Soc.* **240**, 195-218 (1989).
17. Dunn, A. M. & Laflamme, R. *Mon. Not. R. astr. Soc.* **264**, 865-874 (1993).
18. Lynden-Bell, D., Lahav, O. & Burstein, D. *Mon. Not. R. astr. Soc.* **241**, 325-345 (1989).
19. Strauss, M. A., Yahil, A., Davis, M., Huchra, J. P. & Fisher, K. *Astrophys. J.* **397**, 395-419 (1992).
20. Sandage, A. *Astrophys. J.* **307**, 1-19 (1986).
21. Buta, R. J. & McCall, M. J. *Mon. Not. R. astr. Soc.* **205**, 131-152 (1983).
22. Luppino, G. A. & Tonry, J. L. *Astrophys. J.* **410**, 81-86 (1993).
23. Van den Bergh, S. *Astr. J.* **107**, 1328-1331 (1994).
24. Spinrad, H. et al. *Astrophys. J.* **163**, L25-L31 (1971).
25. McCall, M. L. *Astr. J.* **97**, 1341-1349 (1993).

ACKNOWLEDGEMENTS. We thank H. Kahlmann and A. Foley for fast scheduling of WSRT, A. Foley and M. Verheijen for reducing the WSRT image, R. Assendorp for an illustrative co-added 60 μ m IRAS image, R. Tilanus, C. Aspin & T. Geballe for obtaining the UKIRT K-band image, S. Hughes and S. Maddox for their help with the INT-images, M. Irwin and M. Balcells with the optical spectroscopy from the WHT, and N. Brosch, E. Goldberg and E. Almozino for communicating to us the follow-up observations taken at the Wise Observatory. We are grateful to G. Hau for his contribution in early stages of the DOGS project and for discussions on the appearance of Dw1 on the Palomar plate. The assistance of D. Moorrees with the 25-m observing is especially appreciated. The Dwingeloo 25-m radio-telescope and the WSRT are supported by the Netherlands Foundation for Scientific Research (NWO).

Photothermal spectroscopy with femtojoule sensitivity using a micromechanical device

J. R. Barnes*, R. J. Stephenson*, M. E. Welland*, Ch. Gerber† & J. K. Gimzewski†

* Department of Engineering, Cambridge University, Trumpington Street, Cambridge CB2 1PZ, UK

† IBM Research Division, Zürich Research Laboratory, CH-8803 Rüschlikon, Switzerland

WHEN a material absorbs a photon, a fraction of the energy may be transformed into heat. A measurement of photothermal heating as a function of wavelength can provide an absorption spectrum of the material. We have recently^{1,2} developed a micromechanical sensor capable of detecting heat changes of the order of picojoules (10^{-12} J). The instrument incorporates a bilayer cantilever of micrometre dimensions which bends in response to heating. Here we show that this device can be used for photothermal spectroscopy with a power sensitivity of 100 pW—two orders of magnitude better than the sensitivity of conventional photothermal deflection spectroscopy³. The small size of the sensor allows picogram quanti-

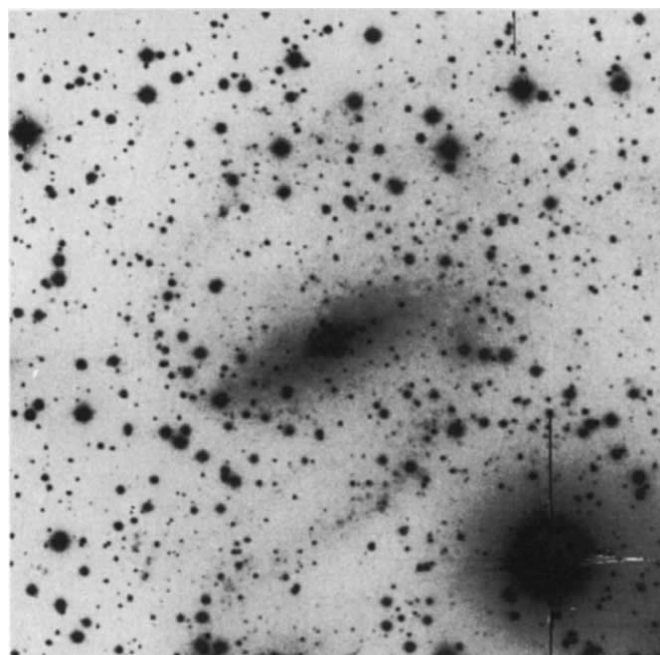


FIG. 3 Composite V, R, I image (produced with the help of S. Hughes and S. Maddox, Cambridge) based on observations at the Isaac Newton telescope. The displayed 484×484 pixels of 0.6 arcsec cover an area of 4.8×4.8 arcmin. The large diameter visible on this image is ~ 4.2 arcmin. Dw1 (centre of image) has a distinct bar, with two spiral arms that can be traced over nearly 180° . The morphology in this figure agrees with that of an SBb galaxy. However, because of the foreground obscuration we could be seeing only the inner part of a much larger SBb or even SBc galaxy.

ties of material to be studied, opening up the possibility of spectroscopic studies on individual cells and bacteria. Being based on silicon technology, the sensor should be compatible with micro-electronic circuitry.

Photothermal deflection spectroscopy involves indirect measurement of the thermal heating induced by photon absorption. This heating is detected by probing the changes in refractive index of a liquid medium in close proximity to the sample surface³. Other more direct (but less sensitive) methods use the pyroelectric effect to measure heat generation at the surface of poly(vinylidene fluoride) films⁴. Both methods use milliwatt-level optical power to produce a measurable effect. Our new spectroscopic method is based on a simple micromechanical sensor which has recently been shown^{1,2} to function as a highly sensitive calorimeter. The basis of our photothermal device is a 'bimetal-

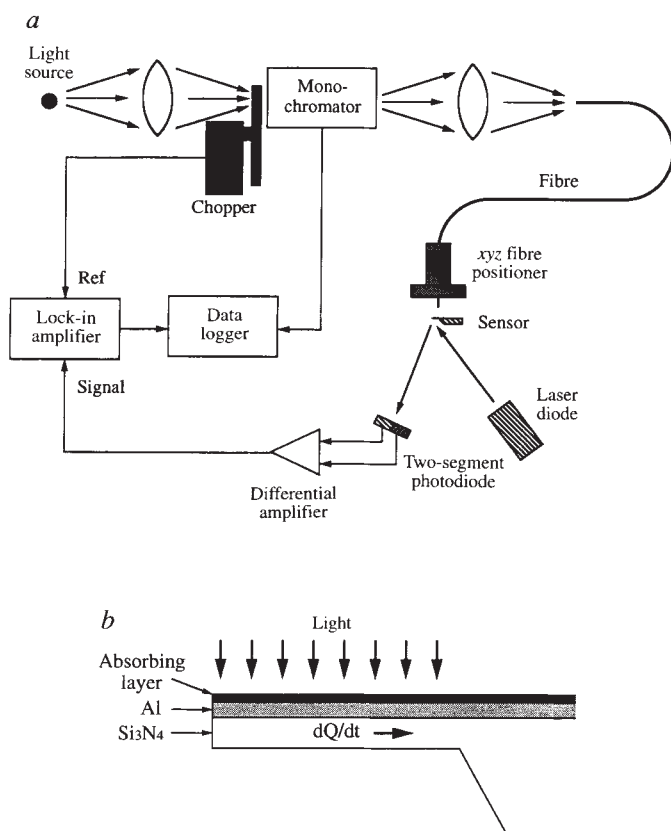


FIG. 1 *a*, Experimental arrangement for measuring photothermal absorption spectra. Light from a 40 W wideband tungsten-halogen source is mechanically chopped at 73 Hz before passing through a monochromator and coupling into a 200- μ m-diameter fused-silica fibre. This is positioned above the micromechanical sensor. Heat absorbed by the sample on the sensor increases the temperature of the sensor causing it to bend because of a 'bimetallic' effect. This bending is detected by an optical deflection system where the movement is measured by sensing the position of a laser spot reflected off the sensor onto a two-segment photodiode. The amplitude of the sensor movement is measured by lock-in amplification using the chopper signal as the reference. We used a modulation frequency of 73 Hz; at this frequency the sensor has a sufficiently fast thermal response to follow the square-wave optical modulation. *b*, Under steady-state conditions there is a constant heat flux, dQ/dt , along the cantilever producing a static sensor deflection. If the modulation period ($1/73$ s) is much greater than the thermal response time constant of the sensor (0.45 ms), as is the case here, then the sensor behaves in a quasi-static fashion and the maximum sensor deflection corresponds to the static deflection value.

lic' cantilever constructed from silicon nitride (Si_3N_4) with dimensions 200 μ m long, 0.6 μ m thick and 40 μ m wide, coated with a 50-nm-thick layer of aluminium (Al). Heat produced during optical absorption in a thin sample deposited on the Al surface of the cantilever induces bending, due to the differing thermal expansion coefficients of Al and Si_3N_4 . The sensor was calibrated by using a He-Ne laser to heat the Al layer^{5,6}, giving a sensitivity of 6 pW nW^{-1} .

The experimental arrangement is shown in Fig. 1. Thermally induced bending of the cantilever is measured by monitoring the position of a laser beam reflected off the underside of the sensor⁷. The bending as a function of wavelength generates a photothermal spectrum. All measurements were done at atmospheric pressure although the device can also be operated in vacuum or liquid environments.

We first demonstrate a photothermal spectrum from a thin film of hydrogen-terminated amorphous silicon (a-Si:H_x); this material is commonly used as a photodetector and in thin-film-transistor flat panel displays. The optical absorption spectrum of a-Si:H_x gives a means of measuring the pseudo-bandgap⁸ which can vary widely, depending on the growth conditions of the film. It is important, for technical applications, to be able to measure this pseudo-bandgap both accurately and simply. The sample consists of a 50-nm-thick film with a total mass of only 0.15 ng. Figure 2 shows the spectrum obtained by scanning over a photon energy range of 2.76–1.77 eV (450–700 nm). The broad-band absorption feature is typical of the indistinct band-edge of a-Si:H_x and shows good agreement with photothermal deflection spectroscopy results⁹. In a typical photothermal deflection spectroscopy measurement, we note that the minimum detectable power is ~ 10 nW (ref. 3). In contrast, preliminary

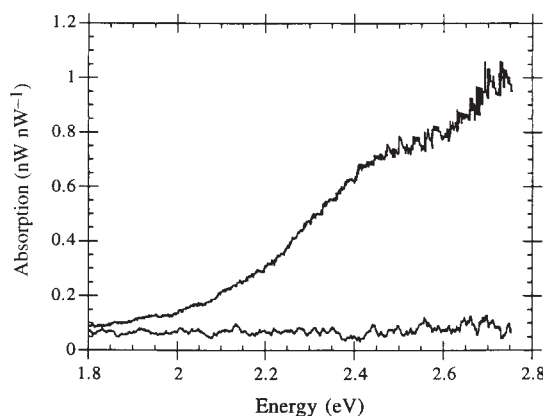


FIG. 2 Top trace, photothermal absorption spectra for a 50-nm film of intrinsic a-Si:H_x . The sensor was completely illuminated by the source (with a fraction of the source illumination actually striking the sensor) while source power varied from 18 nW to 2 nW over the range of wavelengths. The spectrum has been normalized to correct for both of these effects and an increase in the noise is seen at higher photon energies because of this fall in illuminating power. Measurement was made with an noise equivalent bandwidth of 8 mHz over a period of ~ 40 min. The sensor has a measured noise equivalent power of 1.1 nW $\text{Hz}^{-1/2}$, corresponding to a sensitivity of 100 pW for an 8 mHz bandwidth. The absorption spectra shows a wide absorption band-edge typical of a-Si:H_x . The uncoated detector response (bottom trace) is shown for comparison. Here absorption is by the Al layer of the sensor and corresponds to a power of ~ 1.8 nW for a normalized source power of 18 nW. This absorption by the Al layer sets a minimum value for the absorbed power at $\sim 10\%$ of the incident power for optical wavelengths, so that the minimum detectable power equals the measured sensitivity (100 pW) with an incident optical power of ~ 1 nW.

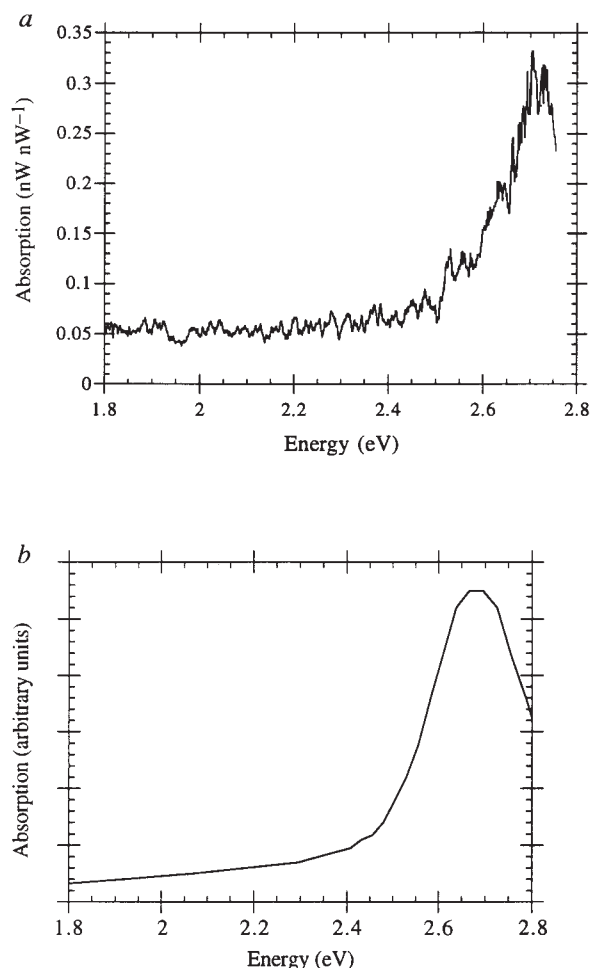


FIG. 3 *a*, Photothermal absorption spectra for fluorescent fluorescein dye molecules fixed in the matrix of 140-nm-diameter latex spheres deposited as a monolayer on the sensor. Measurement conditions are identical to those used to obtain the data in Fig. 2. An absorption peak at 2.7 eV (460 nm) is seen which correlates well with the expected peak; see below. *b*, Absorption spectrum of latex spheres (from *a*) in distilled water, measured with a Philips PU8730 ultraviolet/visible spectrophotometer with 2-nm fixed bandwidth. The absorption peaks at 2.68 eV.

experiments on our device show a sensitivity of 100 pW, representing an improvement by a factor of 100 over photothermal deflection spectroscopy. From these data and the measured thermal time constant of the sensor (0.45 ms), the heat-flux sensitivity of our present device has been estimated at 150 fJ. In addition, it is important to note that quantitative determination of the power (or heat) absorbed is straightforward and direct, compared to photothermal deflection spectroscopy. This is a consequence of the simple operating principle.

Whereas the electronic structure of α -Si:H_x gives a broad absorption spectrum, many organic molecules exhibit relatively narrow absorption lines. We therefore also studied a fluorescent molecule (fluorescein based) which has a narrow absorption band at 2.65 eV (469 nm) and emission at 2.44 eV (509 nm). The difference between these energies results in heat generation when this compound is illuminated with light at the absorption wavelength. A sample of 140-nm-diameter latex spheres, with fluorescein dye molecules in the latex matrix (Duke Scientific Corp., Palo Alto, California), was deposited on the sensor: the total dye mass was of the order of tens of picograms. The photothermal

spectrum (Fig. 3*a*) of this sample displays a single sharp peak at 2.7 eV (460 nm), in excellent agreement with the absorption energy observed for a bulk sample measured with a Philips PU8730 spectrophotometer (Fig. 3*b*). The amplitude and shape of the photothermal response remained unchanged over the timescale of several repeated measurements. This indicates that photothermal bleaching was not observed, as expected from the low optical power of the incident illumination.

The ultimate sensitivity of the device can be calculated by assuming that thermal noise in the cantilever is the limiting factor. Below the mechanical resonance frequency of the cantilever, this amounts to 1.4 pm Hz^{-1/2} (ref. 10). At a noise equivalent bandwidth of 8 mHz the minimum detectable power is 10 pW in our present device, corresponding to a theoretical value of 25 fJ for the minimum detectable heat flux.

This fundamental sensitivity of the present device offers new possibilities for the investigation of a wide range of materials. In addition to its demonstrated performance as a new method for performing photothermal spectroscopy, it has several inherent features giving it potential beyond that of conventional photothermal spectroscopy. Perhaps the most important is the combination of sensitivity and detector size. Although other types of detector have a similar energy sensitivity per unit area¹¹, our device, being micromechanical in nature, has a significantly smaller area (<10⁻⁸ m²), making it possible to measure energy absorption processes in single biological cells. Also, the fact that the sensor works in an aqueous environment makes *in vitro* studies of immobilized cells possible. For example, a single photosynthesizing chloroplast with an energy conversion efficiency of 50% can produce 10–50 pW of heat (ref. 12 and T. ap Rees, personal communication); a metabolizing green alga cell has a heat output of 1–8 pW (ref. 13). The former is within the sensitivity of the present detector, and the latter could be detectable by optimizing the geometry of the lever. An extension of such measurements could allow for fabrication of a whole-cell immunobiosensor, where the heat generated by cellular metabolic processes associated with the cell-antigen reaction could be used as a diagnostic technique. Finally, because of the sensor's origins in force microscopy, it is possible to integrate the sensor with an imaging tip in the manner of scanning near-field optical microscopy¹⁴ so as to combine spatial and energy resolution. □

Received 6 June; accepted 6 October 1994.

- Gimzewski, J. K., Gerber, Ch., Meyer, E. & Schlittler, R. R. *Chem. Phys. Lett.* **217**, 589–594 (1994).
- Meyer, E., Gimzewski, J. K., Gerber, Ch. & Schlittler, R. R. in *The Ultimate Limits of Fabrication and Measurement* (eds Gimzewski, J. K. & Welland, M. E.) (NATO ARW, Kluwer, Dordrecht, in the press).
- Amer, N. M. & Jackson, W. B. in *Semiconductors and Semimetals* Vol. 21 (ed. Pankove, J. K.) 83–112 (Academic, Orlando, 1984).
- Coufal, H. *Appl. Phys. Lett.* **44**, 59–61 (1984).
- Marti, O. et al. *Ultramicroscopy* **42–44**, 345–350 (1992).
- Allegrini, M. et al. *Ultramicroscopy* **42–44**, 371–378 (1992).
- Meyer, G. & Amer, N. M. *Appl. Phys. Lett.* **53**, 1045–1047 (1988).
- Street, R. A. *Hydrogenated Amorphous Silicon* (Cambridge Univ. Press, 1991).
- Fournier, D. & Claude-Boccar, A. in *Photoacoustic and Thermal Wave Phenomena in Semiconductors* 237–255 (ed. Mandelis, A.) (Elsevier, New York, 1987).
- Heer, C. V. *Statistical Mechanics, Kinetic Theory and Stochastic Processes* (Academic, Orlando, 1972).
- Stuckless, T., Al-Sarraf, N., Wartnaby, C. & King, D. A. *J. chem. Phys.* **99**, 2202–2212 (1993).
- Gust, D., Moore, T. A. & Moore, A. L. *Accs chem. Res.* **26**, 198–205 (1993).
- James, A. M. (ed.) *Thermal and Energetic Studies of Cellular Biological Systems* (Inst. of Physics, Bristol, 1987).
- Pohl, D. *Adv. opt. Electron Microsc.* **12**, 243–312 (1991).

ACKNOWLEDGEMENTS. We thank P. Morgan for depositing the amorphous Si samples, S. O'Shea, T. Rayment and C. Woodburn for their assistance with the experiment, and D. Cebon for useful discussions on the bimetallic bending of beams. R. J. S. acknowledges the support of a Canadian NSERC scholarship and a Sir Winston Churchill scholarship.

Rapid climate transitions in a coupled ocean-atmosphere model

Stefan Rahmstorf

Institut für Meereskunde, Düsternbrooker Weg 20, 24105 Kiel, Germany

RECENT geochemical data^{1,2} have challenged the view that rapid climate fluctuations in the North Atlantic at the end of the last glacial were caused by the thermohaline circulation of the ocean being switched 'on' or 'off'. Instead, these data suggest that the circulation pattern must have switched between a warm, deep mode and a cold, shallow mode, probably associated with different sites of deep convection³. Here I present simulations with a three-dimensional ocean model, coupled to an idealized atmosphere, which show this kind of transition. The mechanism for the transition is a rearrangement of convection in the North Atlantic, triggered by a brief freshwater pulse. This results in a drop in sea surface temperature in the North Atlantic by up to 5 °C within less than 10 years. The rate of North Atlantic Deep Water (NADW) formation is the same in the cold climate as in the warm climate, but NADW sinks to intermediate depths only, while Antarctic Bottom Water pushes north to fill the entire abyssal Atlantic.

Palaeoclimate studies show that rapid climate changes in the North Atlantic region are correlated with changes in NADW flow^{2,4,5}. It is widely believed that because the NADW thermohaline 'conveyor belt' has the capacity to transport large amounts of heat towards the northern North Atlantic (at present $\sim 0.5 \times 10^{15}$ W, leading to a warming of 4 °C compared with the northern North Pacific^{6,7}), some of the climate changes visible in the palaeoclimate records are caused by variations in the conveyor belt. This idea is supported by modelling studies, which have demonstrated that the conveyor belt is a self-sustaining circulation which can collapse when disturbed, and which can have multiple steady states⁸⁻¹⁰. Within this picture, a large freshwater perturbation, like the sudden inflow of melt water just before the Younger Dryas event, could have triggered the complete collapse of the conveyor belt, leading to a rapid cooling of the North Atlantic by several degrees within about a decade.

More recent data do not fit this interpretation, however^{1,2}. They show that even during cold periods the conveyor did not stop; NADW flow continued at similar intensity. However, the sites of deep convection appear to have shifted, and NADW flow was probably shallower than during warm periods. Previous model studies of the conveyor are unable to explain these results.

The experiments reported here, using the Geophysical Fluid Dynamics Laboratory (GFDL) ocean circulation model, are consistent with the new data. I have used an idealized configuration with two square ocean basins of equal size, representing Pacific and Atlantic, (as in ref. 10; see Fig. 3a). The model was driven by a simple zonal-mean wind stress. The freshwater flux at the ocean surface was diagnosed from a spin-up experiment, zonally averaged and then held fixed. The traditional 'restoring' boundary condition¹¹ on temperature, used in previous ocean model studies, was replaced by

$$Q = \gamma(T^* - T_o) - \mu \nabla^2(T^* - T_o) \quad (1)$$

This equation derives from an energy balance model of the atmosphere with diffusive horizontal transport of temperature anomalies¹². Q is the heat flux at the ocean surface, T_o the ocean surface temperature, γ a radiative relaxation constant ($3 \text{ W m}^{-2} \text{ K}^{-1}$), μ a constant related to atmospheric heat diffusion ($8 \times 10^{13} \text{ W K}^{-1}$), and T^* defines a climate without oceanic heat transport, around which equation (1) is linearized. T^* was prescribed as a cosine function of latitude, symmetric in both

hemispheres. In contrast to classical restoring, where the atmospheric temperature is assumed fixed and the heat flux from the ocean disappears into an 'infinite sink' atmosphere, equation (1) represents a closed heat budget where heat leaving the ocean can only be lost to space through long-wavelength radiation (first term) or redistributed laterally within the atmosphere by turbulent diffusion (second term). Atmospheric temperature is a diagnostic variable in this model; from equation (1) it has already been eliminated by inserting the atmospheric heat budget equation. Equation (1) is thus suitable for climate studies and allows the prediction of changes in surface temperature resulting from ocean circulation changes.

A conveyor-belt circulation was spun up in the usual way^{10,12}. The resulting model equilibrium has a realistic temperature contrast (not shown here) between North Atlantic and North Pacific due to the conveyor's heat transport and in spite of the zonally uniform forcing. Another realistic feature is the appearance of an Antarctic Bottom Water (AABW) cell underneath the NADW cell in the Atlantic (Fig. 1a), caused by colder temperatures in the Southern Ocean compared to the northern convection region. Previous ocean models had to prescribe these colder temperatures in the forcing field in order to get AABW; here they are a natural consequence of the conveyor circulation.

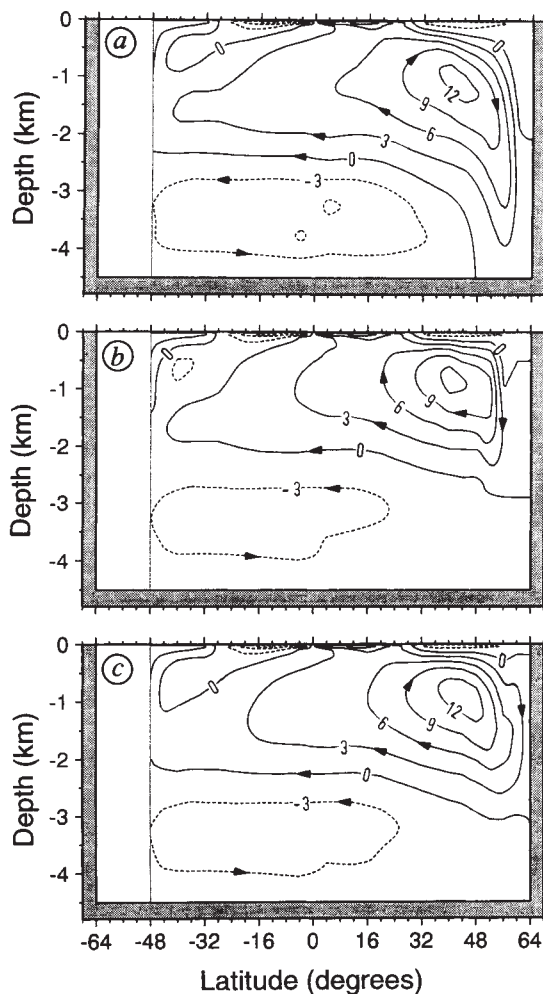


FIG. 1 Zonally integrated meridional mass transport in $10^6 \text{ m}^3 \text{ s}^{-1}$ in the Atlantic model basin: a, for the spin-up conveyor state; b, for the equilibrium reached after the first freshwater perturbation; c, for the equilibrium after the second perturbation. These three coupled equilibria exist under the same forcing. Note that in b and c the North Atlantic Deep Water (NADW) cell does not reach down to the bottom of the North Atlantic.

To test the stability of this state, a strong freshwater pulse was then added to the North Atlantic north of 56°N for 4 years. Time series of important model parameters during this experiment are shown in Fig. 2. The freshwater surge temporarily interrupts convection and slows the conveyor, but 20 years after the end of the perturbation deep water formation has recovered to its original rate. Temperature in the northern North Atlantic region has changed permanently and drastically, however, dropping by an average 2.6°C north of 48°N , and 3°C north of 56°N . Average salinity north of 56°N has dropped by 1‰. At the same time, NADW flow has become shallower, while AABW now penetrates northward in the abyssal Atlantic up to the northern wall (Fig. 1b); similar northward AABW incursions are also found in palaeoclimate data¹³. The pattern of sea surface temperature change is shown in Fig. 3a. Cooling occurs in the whole North Atlantic with a maximum in high latitudes, particularly in the west where the northward flowing boundary current is weakened. These features resemble the pattern of observed interdecadal variability during this century¹⁴. The remainder of the ocean surface warms, most notably in the South Atlantic, where the increase exceeds 0.5°C in some regions.

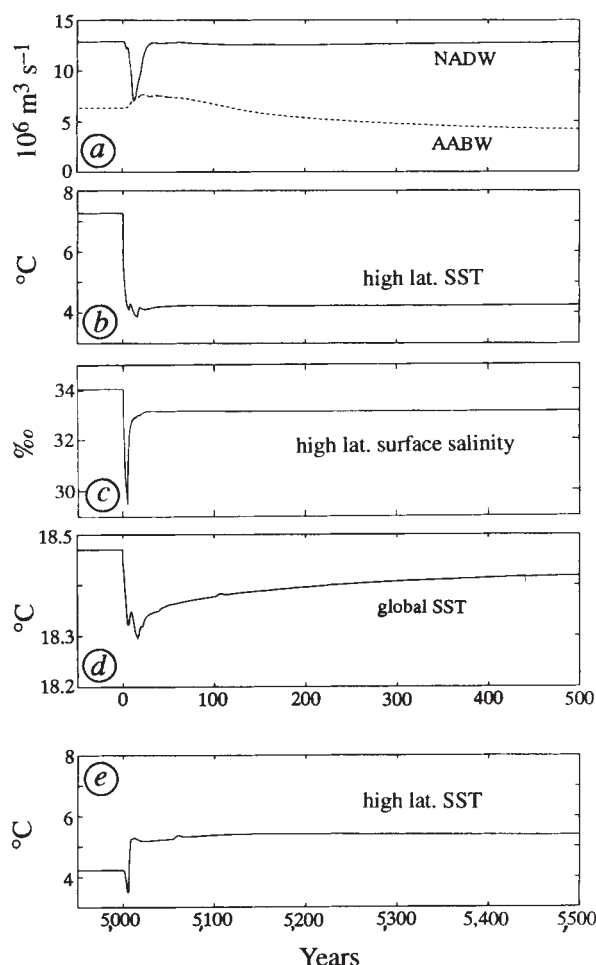


FIG. 2 Time series of model parameters during and after the freshwater perturbation (starting at $t=0$ for 4 years). a, Strength of the NADW cell (solid) and the Antarctic Bottom Water (AABW) cell (dashed) in the Atlantic basin. b, Mean sea surface temperature (SST) in the Atlantic north of 56°N (high lat.). c, Mean sea surface salinity in the Atlantic north of 56°N . d, Mean SST of the entire model domain. e, Same as b, but for the second freshwater perturbation (at $t=5,000$ yr). Note that even the 'global' SST (d) is affected by the perturbation, due to transient heat uptake by the deep ocean, although its equilibrium value is fixed by equation (1) (see text).

The model was integrated for 5,000 years after the perturbation to reach equilibrium, and then the same freshwater perturbation was applied once more. Again a permanent transition to a different circulation state was triggered, this time accompanied by moderate warming in the North Atlantic region (Figs 1c, 2e). The deep-water formation rate in all three equilibria is remarkably similar, probably due to a negative temperature feedback discussed by Rahmstorf¹⁵.

Unlike the 'on/off' transition found in previous models, caused by the advective feedback which sustains the conveyor (as in Stommel's model of two adjacent boxes¹⁶), the new mechanism is due to a convective feedback (described in Welander's model of two vertically stacked boxes¹⁷), which makes convection partly a self-sustaining process. Lenderink and Haarsma¹⁸ have recognised that this convective feedback leads to multiple steady states in three-dimensional circulation models, because some model grid points can have convection 'on' or 'off' under the same forcing, depending only on initial conditions. A large number of model equilibria with different convection patterns are therefore possible. With a realistic atmospheric heat budget like equation (1), the result is different climate states and the possibility of transitions between these states.

Figure 4 shows the convection patterns for the three different model equilibria. The transition between these states was in this case initiated when the freshwater inflow (at a rate of $0.32 \times 10^6 \text{ m}^3 \text{ s}^{-1}$) capped the high latitudes, effectively erasing

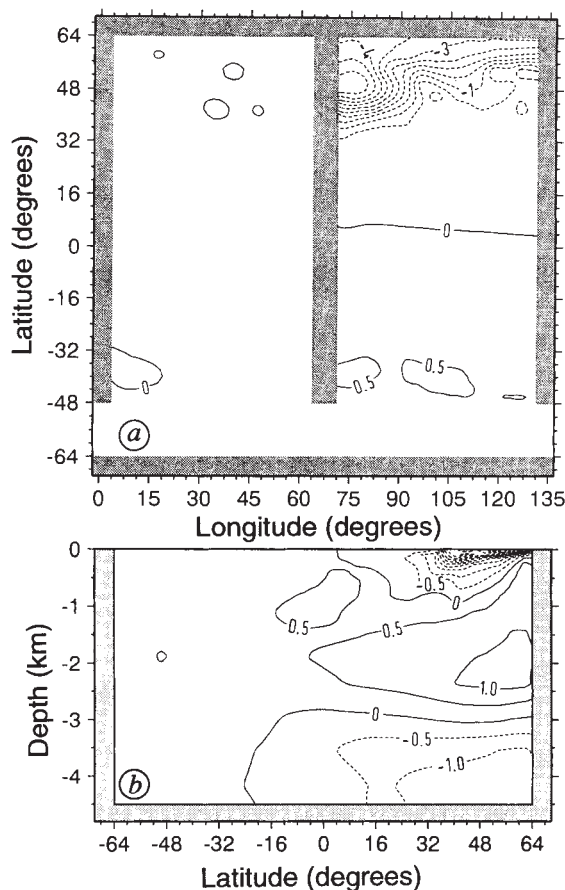


FIG. 3 Spatial structure of the temperature difference between warm and cold conveyor solution, shown in Fig. 1a b. Contour interval, 0.5°C . a, Surface map. The maximum temperature drop is 5.3°C in the western boundary current. b, Latitude–depth section along the western Atlantic basin. Cooling is concentrated near the surface in the North Atlantic, while the NADW outflow level warms. Some cooling occurs below 3 km depth due to the northward incursion of AABW.

the former convection pattern. More controlled transitions without blanket interruption of convection are also possible, using small 'surgical' freshwater perturbations targeted at single convection cells¹⁵. The actual meltwater release during the Younger Dryas event has been estimated as peaking at $0.44 \times 10^6 \text{ m}^3 \text{ s}^{-1}$ (ref. 19).

A simple way of estimating the conveyor's heat transport is to multiply its mass transport by the temperature difference between warm inflow and cold deep outflow (see below). If this heat transport is large, the high latitudes are warm. But a large heat transport (given a fixed mass transport) requires cold outflow, seemingly at odds with warm high latitudes. This apparent contradiction is explained by the fact that convection can con-

nect the deep outflow to the surface in different places, allowing its temperature to increase when the high latitude surface temperature drops. This is clearly seen in Fig. 3b. What has been dubbed a 'cold' conveyor (associated with cold sea surface temperature, SST) is thus one with a warm deep branch and reduced heat transport, due to relatively warm deep-water formation sites. Deep-water temperature and thus meridional heat transport seem to depend mainly on the latitude of convection; when convection moves south after the first perturbation the surface temperature drops, and when convection moves north after the second perturbation it increases again (see Fig. 4).

The difference in oceanic heat transport between the warm conveyor (Fig. 1a) and the cold conveyor (Fig. 1b) peaks at $1.1 \times 10^{14} \text{ W}$ at 44° N , reducing the conveyor's heat transport across this latitude by 30%. In equilibrium this is balanced by a reduction in surface heat loss of 13 W m^{-2} averaged over the region north of 44° N . To achieve this heat flux change solely through longwave radiation (first term in equation (1)) would require an average SST change of 4.2° C . The actual change in SST over this region is only 2.3° C , demonstrating how part of the change in the conveyor's heat transport is compensated by a change in atmospheric transport (second term in equation (1)). The compensation by atmospheric transport is less effective for the change in heat budget than for the conveyor's budget as a whole, because the convective temperature changes are concentrated so far north. If the conveyor's heat transport is written as $c_p \rho m \Delta T$ (where c_p and ρ are heat capacity and density of sea water, m is the conveyor's mass transport of $13 \times 10^6 \text{ m}^3 \text{ s}^{-1}$, and ΔT is a characteristic temperature difference between warm northward- and cool southward-flowing branch), then a change in ΔT of 2° C is sufficient to cause the heat transport change discussed above (compare this to Fig. 3b).

These multiple conveyor-belt states are similar to the ones found recently by Weaver and Hughes²⁰ and linked by these authors to rapid interglacial climate fluctuations. The crucial difference is that climate changes in their model are caused by changes in the conveyor's mass transport rather than its temperature. This model was driven with a traditional restoring boundary condition on SST, corresponding to fixed atmospheric temperature.

Are the climate transitions found in our model realistic? Preliminary results with a realistic topography global ocean model show similar multiple convection states (S.R., manuscript in preparation). Although 'pure' multiple equilibria under the same forcing would never arise in the real ocean under an ever-changing, 'noisy' atmosphere, the main conclusion of this model study remains valid: a shift in convection sites, whether it occurs through atmospheric changes, meltwater inflow or some other cause, could change the conveyor's heat transport and North Atlantic climate, even if deep-water formation continues at a similar rate.

A further question is whether present ocean climate models, which are hydrostatic and parametrize convection as a vertical mixing of statically unstable water, can represent deep-water formation in a realistic way. The crucial issue is whether convection events in nature act as 'mixing agents', causing vertical homogenization of the water column but practically no direct net sinking, or whether they act as a 'conduit' through which water is channelled downward to form new deep water. Send and Marshall²¹ have recently shown that the former is true, justifying the approach taken in ocean circulation models.

Although our idealized model does not address this issue, it is tempting to speculate which regional shifts in convection sites might occur in the real ocean. Present NADW formation sites are the Norwegian Sea and Labrador Sea. Lehman and Keigwin²² have suggested that the conveyor did not cease entirely during cold periods, but that only its lower branch, which originates in the Norwegian Sea, was interrupted. The model results presented here are consistent with this idea and give it an additional twist: they suggest that when deep-water formation is

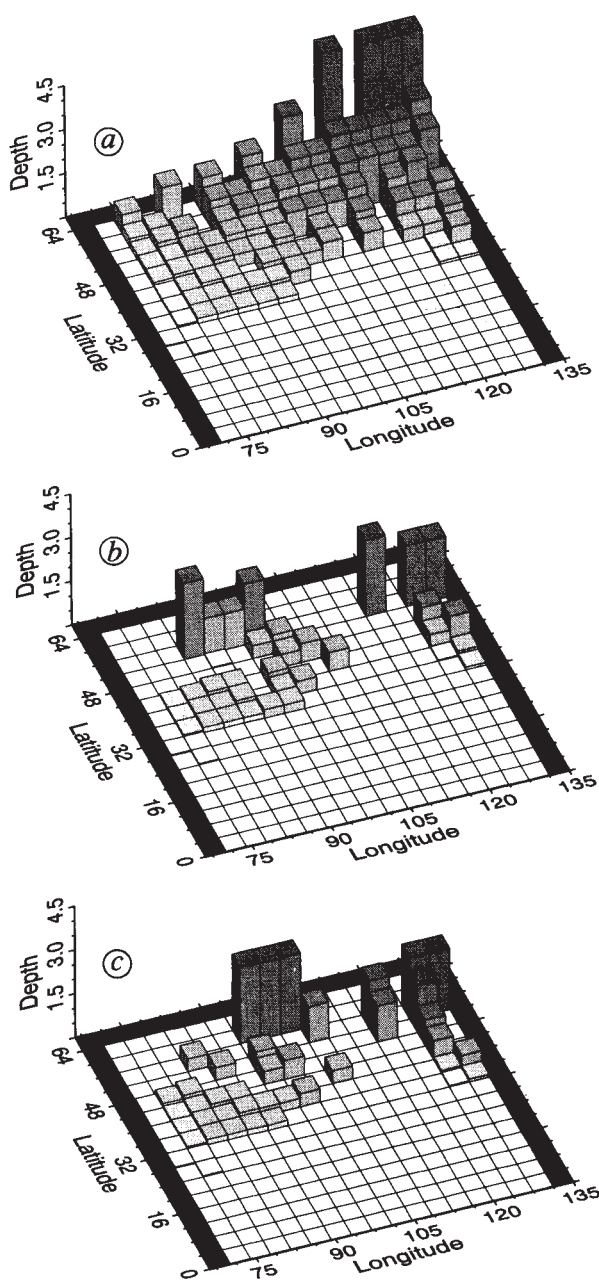


FIG. 4 Convection patterns in the North Atlantic basin, for the three equilibria shown in Fig. 1. The depth (km) of convection in each grid cell is plotted as a vertical bar. Shading is proportional to the number of model levels involved in convection.

reduced in the Norwegian Sea it may increase in the Labrador Sea, so that the two regions act like a see-saw. Climate cooling is then not caused by an overall weakening of the conveyor, but by a reduction in its heat transport related to the higher temperature of Labrador Sea convection.

These model experiments demonstrate that in order to explain rapid cooling events we need not invoke a total collapse of the conveyor; rather more subtle changes can cause a temperature drop in the northern North Atlantic of up to 5 °C. □

Received 16 June; accepted 23 September 1994.

1. Veum, T. et al. *Nature* **356**, 783–785 (1992).
2. Lehman, S. J. & Keigwin, L. D. *Nature* **356**, 757–762 (1992).
3. Zahn, R. *Nature* **356**, 744–746 (1992).
4. Boyle, E. A. & Keigwin, L. *Nature* **330**, 35–40 (1987).
5. Keigwin, L. D. et al. *J. geophys. Res.* **96**, 16811–16826 (1991).
6. Trenberth, K. E. & Solomon, A. *Clim. Dyn.* **10**, 107–134 (1994).

7. Levitus, S. *Climatological Atlas of the World Ocean* (US Dept of Commerce, NOAA, Washington DC, 1982).
8. Manabe, S. & Stouffer, R. J. *J. Clim.* **1**, 841–866 (1988).
9. Maier-Reimer, W. & Mikolajewicz, U. in *Oceanography* (eds Ayala-Castañares, A., Wooster, W. & Yáñez-Arancibia, A.) 87–100 (UNAM, México, 1989).
10. Marotzke, J. & Willebrand, J. *J. phys. Oceanogr.* **21**, 1372–1385 (1991).
11. Haney, R. L. *J. phys. Oceanogr.* **1**, 241–248 (1971).
12. Rahmstorf, S. & Willebrand, J. *J. phys. Oceanogr.* (in the press).
13. Sarinthein, M. et al. *Paleoceanography* **9**, 209–267 (1994).
14. Kushnir, Y. *J. Clim.* **7**, 142–157 (1994).
15. Rahmstorf, S. *J. Clim.* (in the press).
16. Stommel, H. *Tellus* **13**, 224–230 (1961).
17. Welander, P. *Dyn. Atmos. Oceans* **6**, 233–242 (1982).
18. Lenderink, G. & Haarsma, R. J. *J. phys. Oceanogr.* **24**, 1480–1493 (1994).
19. Fairbanks, R. G. *Nature* **342**, 637–642 (1989).
20. Weaver, A. J. & Hughes, T. M. C. *Nature* **367**, 447–450 (1994).
21. Send, U. & Marshall, J. *J. phys. Oceanogr.* (in the press).
22. Lehman, S. J. & Keigwin, L. D. *Nature* **358**, 197–198 (1992).

ACKNOWLEDGEMENTS. I thank J. Willebrand for assistance, R. Zahn, M. Rhein and E. Suess for discussions, and D. Smart for improving the clarity of the language in this Letter. This work was supported by the Deutsche Forschungsgemeinschaft; computations were performed at the German Climate Computer Centre in Hamburg.

Experimental simulations of explosive degassing of magma

H. M. Mader*, Y. Zhang†, J. C. Phillips‡, R. S. J. Sparks‡, B. Sturtevant§ & E. Stolper||

* Institute of Environmental and Biological Sciences, Lancaster University, Lancaster LA1 4YQ, UK

† Department of Geological Sciences, University of Michigan, Ann Arbor, Michigan 48109–1063, USA

‡ Department of Geology, University of Bristol, Bristol BS8 1RJ, UK

§ Graduate Aeronautical Laboratories, || Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA

THE violent release of volatiles in explosive volcanic eruptions is known to cause fragmentation of magma and acceleration of the resulting mixture of gas and pyroclasts to velocities exceeding 100 m s⁻¹ (ref. 1). But the mechanisms underlying bubble nucleation, flow acceleration and fragmentation are complex and poorly understood. To gain insight into these phenomena, we have simulated explosive eruptions using two model systems that generate expansion rates and flow velocities comparable to those observed in erupting volcanos. The key feature of both experiments is the generation of large supersaturations of carbon dioxide in a liquid phase, achieved either by decompressing CO₂-saturated water or by rapid mixing of concentrated K₂CO₃ and HCl solutions. We show that liberation of CO₂ from the aqueous phase is enhanced by violent acceleration of the mixture, which induces strong extensional strain in the developing foam. Fragmentation then occurs when the bubble density and expansion rate are such that the bubble walls rupture. In contrast to conventional models of fragmentation^{1,2}, we find that expansion and acceleration precede—and indeed cause—fragmentation.

The experiments were carried out (Fig. 1) at the University of Bristol and the California Institute of Technology using shock-tube techniques first proposed in this context by Bennett³ and more recently developed in studies of volcanic jets⁴, explosive vaporization⁵ and high-speed dense dusty gases^{6,7}. A crucial characteristic of the natural systems that we have tried to match in our experiments is that the volatile component (largely H₂O and/or CO₂ in real volcanic eruptions) is in a gaseous state after exsolution but the liquid component (in nature, a silicate liquid ± crystals), which makes up most of the mass of the system, is essentially entirely condensed, even after the volatile component has nearly completely exsolved. In these respects the systems in our experiments differ significantly from one-component systems⁵ and two-component systems that completely (or nearly completely) evaporate on decompression.

Design of small-scale experiments to model large-scale volcanic phenomena requires consideration of scaling and dynamic similarity, but no scaling relationships have been established⁸ for rapidly degassing bubbly liquids or the high-velocity high-density gas–particle flows that they generate⁷. Thus it is desirable to conduct simulations at the same velocity and with accelerations and flow densities similar to those in the large-scale setting. Velocities in our experiments approach those of volcanic flows (~100 m s⁻¹)¹, as do the accelerations (~100g, ref. 7; note that gravity (1g) is unimportant in flows that experience such large accelerations). The high velocities and accelerations experienced by these flows mean that inertial rather than viscous forces control the dynamics. The ratio of the test-cell pressure to reservoir pressure in the decompression experiments (Fig. 1a) ranges from 30 to 300, comparable to magmas with a few per cent dissolved water (saturation pressures ~50–100 MPa) that disrupt explosively at a few megapascals (refs 2, 9). Obviously, the length scales of the volcanic system cannot be reproduced in the laboratory. Generally, in the study of bubbly liquids and dusty gases, the diameter of the laboratory flow channel is chosen to be much larger than the smallest bubble or particle class being modelled^{7,10} so that the long-range dynamical interactions between phases that generate a variety of flow scales are free to act.

Bubble nucleation and growth are ultimately the processes that drive the accelerations observed in our experiments, so scaling of our results to nature requires an evaluation of how they may differ in supersaturated magmas. Molar fractions of volatiles in our experiments (up to 0.004 for CO₂ in the depressurization experiments, and as large as 0.1 in the chemical injection experiments) are comparable to those of intermediate to silicic magmas (0.02–0.06 for H₂O)², so the ratio of gas to condensed matter is comparable in the experiments and in nature. However, the diffusivity of CO₂ in H₂O (2×10^{-9} m² s⁻¹ at 20 °C)¹¹ is ~200 times larger than that of H₂O (at concentrations of about 0.03) in magma at 850 °C (~10⁻¹¹ m² s⁻¹)^{12,13}. The diffusively-controlled volume growth rate of bubbles during degassing is proportional to the product of diffusivity and concentration, which ranges from ~10 times larger than in magma in the decompression experiments to locally 200 times in the chemical experiments. The viscosity of water is also many orders of magnitude less than that of magma in explosive eruptions. But numerical calculations (ref. 14 and J. Barclay, D. S. Riley and R.S.J.S., unpublished results) indicate that, even under conditions of explosive eruption, diffusive bubble growth is not retarded by viscous effects unless magma viscosities exceed 10⁸ Pa s. In the fragmentation region, with pressures of a few megapascals, magmas are not fully degassed and a few tenths of a per cent of residual dissolved water are sufficient to keep viscosities at values that would not inhibit explosive expansion

rates². Initial bubble densities measured in the decompression experiments are of the order of 10^9 m^{-3} , whereas in laboratory experiments on volcanic glasses¹⁵ densities of the order of $2 \times 10^{14} \text{ m}^{-3}$ have been reported; observations of pumice^{16,17} suggest nucleation densities of 3×10^{14} to $3 \times 10^{16} \text{ m}^{-3}$. That bubble densities in the present experiments are up to 5 orders of magnitude smaller than in nature and in decompression experiments on magmatic compositions is probably more important than the smaller difference in diffusivities noted above and suggests that natural events could be even more explosive than those observed by us in the laboratory.

Figure 2 shows representative single frames taken from high-speed motion pictures of the two experiments. In the experiment

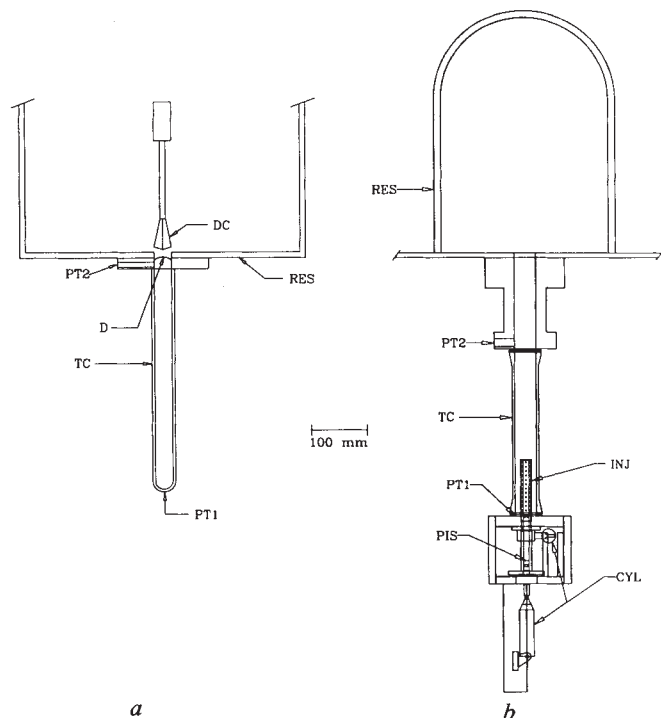


FIG. 1 Schematic diagram of the experimental apparatus. Water saturated with CO_2 is generated in the test cell TC made of Pyrex pipe and erupts into the reservoir RES. Pressures are measured with fast-response piezoelectric transducers at the bottom of the test cell PT1 and at the exit PT2. The flows are visualized by high-speed motion-picture photography at rates up to 6,000 frames per second. *a*, Apparatus at Caltech⁵. The test cell is partially filled with CO_2 -saturated H_2O at pressures up to 0.7 MPa. The reservoir is evacuated to $\sim 7 \text{ kPa}$. Sudden decompression and supersaturation is achieved by puncturing a diaphragm D with a solenoid-driven diaphragm cutter DC. Gas is released from the liquid progressively by bubble nucleation, growth and foam formation. Bubble growth and foam expansion are limited by mass diffusion of solute to the bubble walls. The viscosity of the test mixture was varied from that of water (10^{-3} Pa s) to 0.7 Pa s by the addition of up to 1% organic polymer. *b*, Apparatus at Bristol University. A quantity (4 ml) of K_2CO_3 solution is injected in a few milliseconds through 96 holes in injector INJ into the test cell, which initially contains a much larger volume (100 cm^3) of HCl solution (up to the top of INJ). Mixing of the reactants initially occurs on the boundaries of the 96 jets and later due to stretching of fluid elements as the gas-liquid mixture accelerates up the test cell. The rate of the ensuing chemical reaction is limited by fluid-mechanical mixing and the rate-limiting step $\text{H}_2\text{CO}_3 \rightleftharpoons \text{CO}_2 + \text{H}_2\text{O}$, with reaction timescale (the time for the reactant to reach $1/e$ of its initial concentration) about 70 ms. Production of vapour-phase CO_2 is further limited by mass diffusion. By using high concentrations of the reactants, very large gas yields are achieved with theoretical supersaturation pressures up to several megapascals. The pneumatic cylinders CYL, which drive piston PIS to inject the K_2CO_3 and open the 96 holes, are computer-controlled.

of Fig. 2*a*, velocities up to 14 m s^{-1} and accelerations of more than $200g$ are observed for supersaturation pressures of 0.7 MPa. The flows expand at almost constant acceleration (Fig. 3*a*), which is strongly and positively correlated with the ratio of test-cell pressure to reservoir pressure, and weakly negatively correlated with liquid viscosity. As bubble growth causes expansion and the number of bubbles remains constant following initial nucleation (suggesting heterogeneous nucleation), constant acceleration implies that the bubble volume increases in proportion to t^2 (where t is time), or the bubble diameter increases as $t^{2/3}$. Preliminary measurements on individual bubbles, during the time when bubbles are well-separated and roughly spherical, confirm this relationship. In the flows generated by chemical reaction (Fig. 2*b*), peak velocities and accelerations increase with CO_2 supersaturation (Fig. 3*b*). CO_2 supersaturations up to 9 MPa, velocities up to 30 m s^{-1} and accelerations up to $150g$ are achieved when saturated K_2CO_3 (6 M) is injected into concentrated HCl (12 M). The acceleration of these flows increases approximately linearly with time because

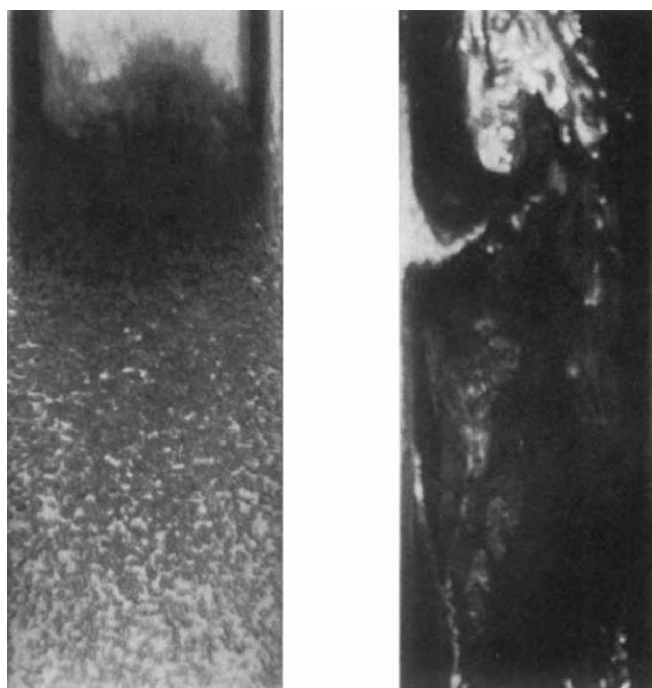


FIG. 2 Photographs of the early development of simulated eruptions. The views are back lit. Light areas are bubble-free liquid or drop-free vapour. A large density of bubbles or droplets scatters light from the beam making the medium opaque (dark). *a*, Flow generated by depressurization. Image of the test cell from 4 to 13 cm above its base. Initial cell pressure, 565 kPa; initial tank pressure, 6.9 kPa. The photograph shows the test solution, initially 10 cm deep, 7.1 ms after depressurization. Bubbles nucleate uniformly ($\sim 10^3 \text{ cm}^{-3}$) throughout the solution immediately on depressurization, and negligible nucleation occurs thereafter. Bubbles near the top of the mixture grow faster than those below. The flow in the photograph has expanded to 12 cm height. Above 10.5 cm the liquid is finely fragmented, while between 9.5 and 10.5 cm the bubbles have merged into a rapidly accelerating foam. The bubbles in the foamed region are stretched in the flow direction. Fragmentation appears to occur by the bursting of bubbles on a very irregular surface, due perhaps to a combination of over-pressurization, rapid stretching and violent interaction with neighbouring bubbles. *b*, Flow generated by chemical reaction. Image of the test cell from 29 cm to 38 cm above its base. The photograph shows the test solution, initially 9.7 cm deep, 34 ms after the initiation of the chemical reaction. Bubble nucleation begins after an incubation time of the order of 5 ms and continues over the timescale of the experiment, as the expanding flow produces strong mixing of the reactants. The resulting foam is highly heterogeneous, with fragmentation occurring in the regions of most violent gas release.

of continuous mixing of the reactants, which is itself enhanced by the stretching of the accelerating two-phase mixture. Eventually the acceleration reaches a maximum and then decreases when the reactants are depleted.

The two sets of experiments reveal several aspects of the physics of violent degassing.

(1) The growth rate measured in the decompression experiments is greater than the $t^{1/2}$ power-law behaviour expected for diffusive growth of a spherical bubble at rest in an infinite fluid and frequently used in models of explosive volcanic eruptions^{2,18}.

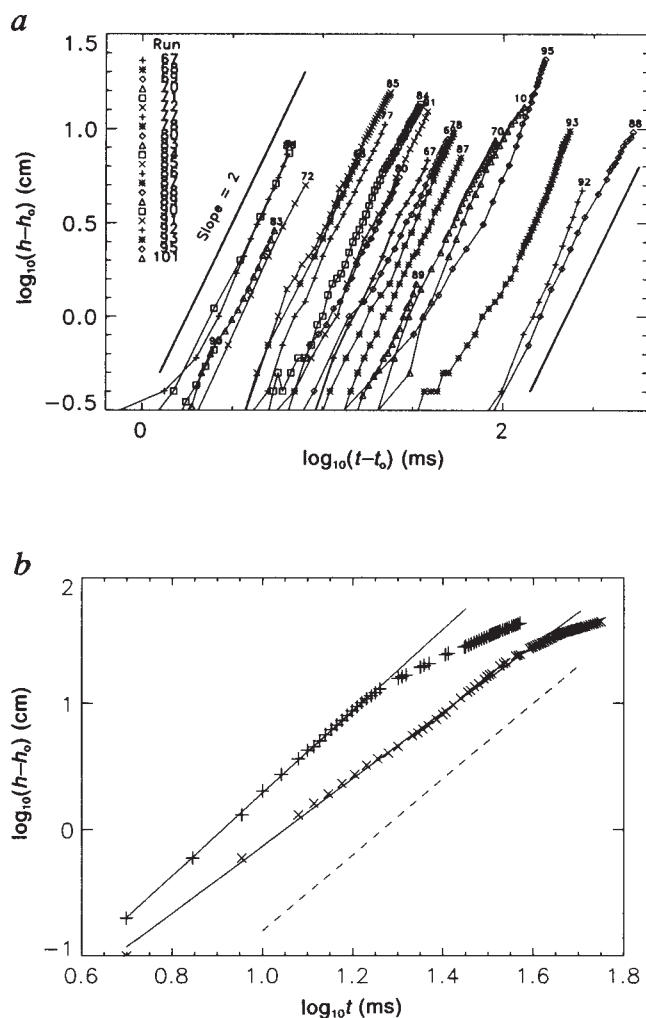


FIG. 3 Position of the flow head h as a function of time t measured from high-speed motion pictures. *a*, Flow generated by depressurization. Log-log (base 10) plot of $h-h_0$ against t . Origin of time, $0 \leq t_0 \leq 22.5$ ms, was adjusted for each run to optimize the linearity of the plot. Results of 22 runs at different saturation pressure, ambient pressure, fluid viscosity and fill depth are shown. Lines of slope 2 (constant acceleration) are indicated at right and left. The acceleration increases monotonically from right to left on the plot from 0.7 m s^{-2} (run 88) to $2,100 \text{ m s}^{-2}$ (run 90). *b*, Flow generated by chemical reaction. Top curve, $6 \text{ M K}_2\text{CO}_3 \rightarrow 12 \text{ M HCl}$; bottom curve, $6 \text{ M K}_2\text{CO}_3 \rightarrow 6 \text{ M HCl}$. Each curve is a composite of data from five films viewing different parts of the test cell. The origin of time for each film was selected to ensure continuity of the plotted data. A least-squares fit (solid lines) of the initial linear part of the data was made to the function $h-h_0 = At^b$ where h_0 is the initial height of the liquid in the test cell, with the following results: $6 \text{ M K}_2\text{CO}_3 \rightarrow 6 \text{ M HCl}$, $b = 2.69$, $A = 1.49 \times 10^{-3}$; $6 \text{ M K}_2\text{CO}_3 \rightarrow 12 \text{ M HCl}$, $b = 3.26$, $A = 1.07 \times 10^{-3}$. As $b > 2$, these flows experience accelerations that increase with time and with CO_2 yield. Line of slope 3 ($a \propto t$) is shown on the right (dashed).

The observed $t^{2/3}$ growth rate has, however, been predicted theoretically¹⁹ for bubbles moving with constant convective velocity U relative to supersaturated liquid. One mechanism for inducing convective motion of bubbles in these experiments and in erupting magmas is buoyant rise in the rapidly-accelerating fluid. Constant-velocity (that is, independent of bubble diameter) buoyant rise occurs in a regime of bubble motion dominated by capillarity and gravity, when $We^3/Fr > 100$ (ref. 20), where $We = \rho U^2 D / \sigma$ is the Weber number, $Fr = U^2 / aD$ is the Froude number, ρ is the liquid density, D is the bubble diameter, σ is the surface tension and a is the acceleration. Under these conditions $U = 1.2(\sigma a / \rho)^{1/4}$. Thus the rise velocity depends weakly on acceleration and surface tension and, for the experiments reported here ($a = 100g$, $\sigma = 0.07 \text{ N m}^{-1}$, $\rho = 1,000 \text{ kg m}^{-3}$) is $\sim 0.6 \text{ m s}^{-1}$; $We^3/Fr = 6,000$ for $D = 2 \times 10^{-3} \text{ m}$. Although bubble rise under different laboratory conditions or in magmatic systems will probably be in a different regime of bubble rise, our experiments demonstrate that in general degassing in explosively accelerating liquids is enhanced by buoyant convective diffusion.

(2) In the chemically-generated flows, large gradients of HCl concentration near the injection jets generate large supersaturation gradients and spatially inhomogeneous liberation of CO_2 vapour. Therefore, the chemically-generated two-phase flows are intrinsically heterogeneous, and thus they may model the behaviour of inhomogeneities occurring in real natural systems. This difference between the two types of flows is visible in Fig. 2; striations of differing opacity, varying from clear (no bubbles) to black (highly vesiculated, fragmented) appear throughout the image of Fig. 2*b*, whereas the dark, high-bubble-density fragmentation region in Fig. 2*a* is confined to the neighbourhood of the free surface. In the chemically-generated flow, even liquid below virtually bubble-free clear fluid (Fig. 2*b*, top) is fragmented. Evidence of heterogeneous vesiculation is common in plinian pyroclastic deposits and ignimbrites.

(3) In our experiments, water fragments into a spray entirely in the liquid state, demonstrating that fragmentation induced by rapid expansion in the liquid state can occur as well in volcanic systems. But in contrast to water, silicic magmas exhibit viscoelastic behaviour²¹ at strain rates comparable to those observed in our experiments; we observe accelerations to 50 m s^{-1} in distances of less than 1 m, so the extensional strain rate is of the order of 100 s^{-1} . In the Maxwell model of linear viscoelastic materials, the relaxation time τ is given approximately by $\tau = \mu / G$, where μ is the shear viscosity and G is the shear modulus²². With $\mu \approx 10^8 \text{ Pa s}$ for magma with a few weight per cent H_2O at eruption temperature² and $G \approx 10^{10} \text{ Pa}$, then $\tau \approx 10^{-2} \text{ s}$, of the same order as the inverse strain rate in rapid exsolution. Thus rapid expansion sufficient to cause magma to cross the glass transition, leading to brittle fracture of the foam, may be possible. Future experiments will examine this mechanism in model viscoelastic fluids.

(4) An important result of our experiments is that models^{1,2} postulating that magma fragments at a simple downward-propagating front in almost static foams with vesicularities of 0.7–0.8, and that flow expansion occurs after fragmentation, should be revised. Foam acceleration precedes fragmentation, rather than the reverse. In addition, in some of the depressurization experiments, vesicularities in excess of 0.7–0.8 are achieved without fragmentation. Silicic pumices with high vesicularities are also known²³.

Convectively-enhanced degassing and the mechanisms of flow expansion and fragmentation will, when incorporated into theoretical models of eruption dynamics, significantly affect such parameters as flow velocity and density at the conduit exit, and the duration of eruptions. Consequently, new insight gained from these experimental simulations could have an important impact on the theoretical prediction of volcanic phenomena and hazards. □

Received 28 March; accepted 28 September 1994.

1. Wilson, L., Sparks, R. S. J. & Walker, G. P. L. *Geophys. J. R. astr. Soc.* **63**, 117–148 (1980).
2. Sparks, R. S. J. *J. Volcan. geotherm. Res.* **3**, 1–37 (1978).
3. Bennett, F. D. *Nature* **234**, 538–539 (1971).
4. Kieffer, S. W. & Sturtevant, B. J. *Geophys. Res.* **89**, 8253–8268 (1984).
5. Hill, L. G. & Sturtevant, B. in *Adiabatic Waves in Liquid–Vapor Systems* (eds Meier, G. E. A. & Thompson, P. A.) 25–37 (Springer, Berlin, 1990).
6. Anilkumar, A. V. thesis, California Inst. Technol. (1989).
7. Anilkumar, A. V., Sparks, R. S. J. & Sturtevant, B. J. *Volcan. geotherm. Res.* **56**, 145–160 (1993).
8. Barnea, D. & Taitel, Y. in *Encyclopedia of Fluid Mechanics* (ed. Cheremisinoff, N. P.) Ch. 16 (Gulf, Houston, 1985).
9. Wilson, L. J. *Volcan. geotherm. Res.* **8**, 297–313 (1980).
10. Sangani, A. S. in *Encyclopedia of Fluid Mechanics* (ed. Cheremisinoff, N. P.) Ch. 5 (Gulf, Houston, 1985).
11. Cussler, E. L. *Diffusion: Mass Transfer in Fluid Systems* (Cambridge Univ. Press, 1984).
12. Zhang, Y., Stolper, E. M. & Wasserburg, G. J. *Geochim. cosmochim. Acta*, **1**, 1–16 (1990).
13. Zhang, Y., Stolper, E. M. & Wasserburg, G. J. *Earth planet. Sci. Lett.* **103**, 228–240 (1990).
14. Dobran, F. J. *Volcan. geotherm. Res.* **49**, 285–311 (1992).
15. Navon, O. & Hurwitz, S. *Terra abstr.* **5**, 574 (1993).
16. Sparks, R. S. J. & Brazier, S. *Nature* **295**, 218–220 (1982).
17. Whitham, A. G. & Sparks, R. S. J. *Bull. volcan.* **48**, 209–223 (1986).
18. Toramaru, A. J. *Geophys. Res.* **94**, 17523–17542 (1989).
19. van Wijngaarden, L. *Int. J. Heat Mass Transfer* **10**, 127–134 (1967).
20. Peebles, F. N. & Garber, H. J. *Chem. Engng. Prog.* **49**, 88–97 (1953).
21. Dingwell, D. B. & Webb, S. L. *Phys. Chem. Miner.* **16**, 508–516 (1989).
22. Scherer, G. W. *Relaxation in Glass and Composites* (Wiley, New York, 1986).
23. Klug, C. & Cashman, K. V. (abstr.) *EOS* **72**, 312 (1991).

ACKNOWLEDGEMENTS. This work was supported by the UK National Environmental Research Council, the BP Venture Fund, the US NSF, the California Institute of Technology and the University of Michigan.

Optimal harvesting, economic discounting and extinction risk in fluctuating populations

Russell Lande*, Steinar Engen†
& Bernt-Erik Saether‡

* Department of Biology, University of Oregon, Eugene, Oregon 97403-1210, USA

† Department of Mathematics and Statistics, University of Trondheim AVH, N-7055 Dragvoll, Norway

‡ Norwegian Institute for Nature Research, Tungasletta 2, N-7004 Trondheim, Norway

DETERMINISTIC models demonstrate that when the economic discount rate of future harvests exceeds a critical value related to population growth rate, the strategy that maximizes the present value of cumulative harvest is immediate extinction (liquidation) of the population^{1–3}. Here we analyse stochastic models to derive optimal strategies that maximize the expected present value of cumulative harvest before extinction of a fluctuating population. Stochastic models reveal that discount rates below the critical value can substantially reduce the mean time to extinction and the expected real harvest before extinction. With an unstable equilibrium at small population size (Allee effect^{4–6} or depensation⁷), the critical discount rate is lower in the stochastic model than in the corresponding deterministic model. These results argue against economic discounting in the development of optimal strategies for sustainable use of biological resources.

Commercially important species are often overharvested to economic depletion (many fisheries and forest-dwelling species^{7,8}), to near extinction (the blue whale, right whale, northern elephant seal, American bison, black rhino, white rhino, for example^{9–13}), or to extinction (Stellar's sea cow, great auk, Caribbean monk seal^{10,14}). Overexploitation, usually combined with habitat destruction and/or introduced species, threatens about one-third of the endangered mammals and birds of the world¹⁰.

Optimal harvesting policies that maximize profit may fail to achieve sustainable use of renewable resources. Economists note that entrepreneurs typically discount future income at a rate equal to or greater than the real interest rate (monetary interest rate minus inflation rate)^{1,2}. When the discount rate, δ , exceeds a critical value, δ^* , deterministic models of population growth show that the economically optimal strategy is immediate harvesting to extinction^{1–3} (Fig. 1). When the discount rate is below the critical value, deterministic models predict a perpetual constant harvest.

Real populations fluctuate because of demographic and environmental stochasticity (individual and temporal variation) in birth and death rates, and palaeontology indicates that extinction is the eventual fate of all species^{15,16}. Stochastic models demonstrate that population fluctuations reduce the maximum expected harvest below the deterministic maximum sustained yield (MSY)^{17–19}; this and the collapse of many fisheries discredited the MSY concept⁷. However, economic discounting and the dynamics of extinction have received little attention in stochastic harvesting models.

Denote the mean annual change in a population of size N as $M(N) = M_0(N) - y(N)$, where $M_0(N)$ represents the expected dynamics without harvesting and $-y(N)$ is the reduction from harvesting. The variance of annual change in population, given the present size, is $V(N)$. The discount rate and the profit per unit harvest are assumed to be constants. If N_t is the population size at time t , then from an initial population N_0 the expected present value of cumulative harvest before extinction is

$$\psi(N_0, \delta) = E \left[\int_0^{t^*} e^{-\delta t} y(N_t) dt \right] \quad (1)$$

where t^* is the extinction time of a particular population in a stochastic environment. Assuming that proportional changes in population per year are usually small, as for large-bodied birds

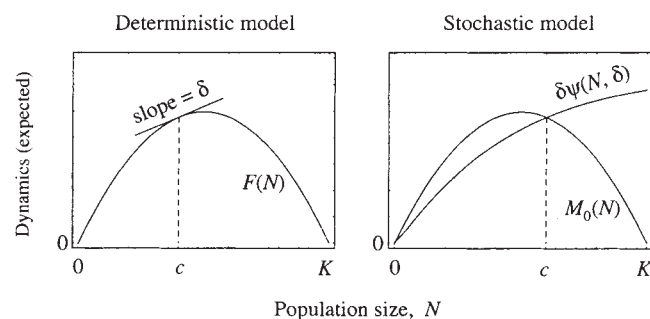
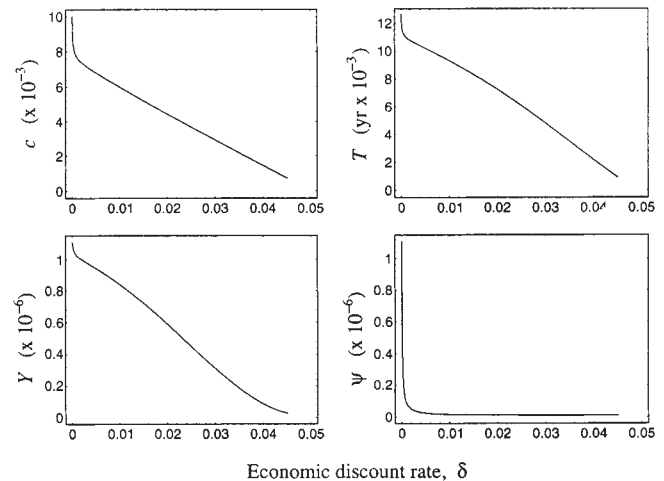


FIG. 1 Optimal harvesting strategies in deterministic and stochastic population models, assuming a constant profit per unit harvest. In the deterministic model, the population dynamics in the absence of harvesting are $dN/dt = F(N)$. The population size that maximizes the present value of future harvests, c , occurs where the slope of the population growth curve equals the discount rate^{1,2}, $F'(c) = \delta$. There may be zero, one or more solutions; the optimal strategy is the solution with the largest sustained yield. When the discount rate exceeds a critical value, δ^* , equal to the maximum $F'(N)$ with $F(N) > 0$, there is no solution and the optimal strategy is immediate harvesting to extinction. The stochastic model has expected dynamics $M_0(N)$ without harvesting. The optimal harvesting threshold, c , above which all excess individuals are immediately harvested and below which no harvesting occurs, maximizes the expected present value of cumulative yield before extinction, ψ , defined in equation (1). c is approximately a solution of $M_0(c) = \delta \psi(c, \delta)$. There may be zero, one, or more solutions. The optimal strategy corresponds either to the solution producing the largest ψ or to $c = 0$. If there is no solution, or if $c = 0$ gives a larger ψ , then the optimal strategy is immediate harvesting to extinction. The (expected) population dynamics illustrated for both the deterministic and stochastic models are given in Fig. 2, and the values of c shown are for $\delta = 0.01$ per year.

FIG. 2 Various quantities in an optimized stochastic harvesting model as a function of the economic discount rate, δ . c is the optimal harvesting threshold that maximizes the expected present value of cumulative harvest before extinction, ψ . Also shown are the mean time to extinction, T , the expected real harvest (not discounted) before extinction, Y , and ψ , corresponding to the optimum threshold, as functions of δ . Without harvesting, the population has expected dynamics $M_0(N) = r(N-A)(1-N/K)$, with an unstable equilibrium at A (from an Allee effect⁴⁻⁶ or depensation², due for example to the difficulty of finding a mate in a sparsely distributed population) and a stable equilibrium at $K > A$. In this model $M_0(N)_{\max} = (1-A/K)r$. Parameters are $r = 0.05$ per year, $K = 10,000$, $A = 20$, and initial population $N_0 = K$. The annual variance in population growth, $V(N) = v_1 N + v_e N^2$, includes both demographic and environmental stochasticity, with $v_1 = 1.0$ and $v_e = 0.04$. The harvesting rate is $y(N) = \infty$ for $N > c$ and 0 otherwise. For low discount rates there are two possible solutions for the optimal threshold, c , the smaller one giving a minimum of ψ , and the larger one giving a maximum of ψ . For higher discount rates these two solutions converge to finally merge and disappear when $\delta > 0.48$. However, $c = 0$ produces the global maximum of ψ when $\delta > 0.45$, which is below the deterministic prediction of $\delta^* = M_0(N)_{\max} = 0.499$.

METHODS. Numerical values of c were obtained from $M_0(c) = \delta \psi(c, \delta)$ with equation (3), which was derived by converting equation (2) to integral form, $\psi(N_0, \delta) = Y - \delta \int_0^\infty \psi(N, \delta) G(N, N_0) dN$ where $G(N, N_0)$, the Green function of the diffusion process for N , gives the amount of time spent at each population size before extinction²⁰. For $N_0 > c$, with an infinite harvesting rate above the threshold, the population is instantly reduced to c , so that $\psi(N_0, \delta) = N_0 - c + \psi(c, \delta)$. The Taylor series for $\psi(N, \delta)$ around N_0 is then exactly linear; substituting it into the preceding integral and solving for $\psi(N_0, \delta)$ produces equation (3). Approximate values from equation (3) were obtained by numerical integration of Y , T and $\bar{N}$ using formulae in ref. 21. These were checked against numerical



Economic discount rate, δ

and mammals, this quantity obeys²⁰

$$\frac{V(N_0)}{2} \frac{d^2 \psi}{dN_0^2} + M(N_0) \frac{d\psi}{dN_0} - \delta \psi = -y(N_0) \quad (2)$$

Extinction occurs when $N_r = 0$, and for realistic models of population growth, the boundary at $N = \infty$ is inaccessible, giving the boundary conditions $\psi(0, \delta) = \psi(\infty, \delta) = 0$.

With no discounting, the type of strategy that maximizes the cumulative harvest before extinction has harvesting rate equal to zero for populations below a threshold, c , and as large as possible for populations above the threshold²¹. With unlimited harvesting capacity, the optimal threshold is simply equal to the stable equilibrium population size in the absence of harvesting, the carrying capacity, at which $M_0(c) = 0$. This remarkable result applies for any form of population dynamics that can be accurately approximated by a diffusion process²¹.

With economic discounting, and an infinite harvesting rate above the threshold, the solution of equation (2) can be accurately approximated for $N_0 > c$,

$$\psi(N_0, \delta) \simeq \frac{Y + (N_0 - \bar{N})\delta T}{1 + \delta T} \quad (3)$$

Y is the expected real harvest (not discounted) before extinction, T is the mean time to extinction, and $\bar{N}$ is the average population size before extinction^{20,21}. The value of the threshold that maximizes ψ is either a solution of $M_0(c) = \delta \psi(c, \delta)$ or $c = 0$. This generalizes the previous result that with unlimited harvesting capacity, and no discounting, the optimal harvesting threshold is at the carrying capacity²¹. In general, economic discounting reduces the optimal harvesting threshold below the carrying capacity.

Optimal strategies can differ greatly in deterministic and stochastic models with the same expected dynamics (Fig. 1). For a zero discount rate the optimal threshold in the stochastic model of Fig. 2 is $c = K$, whereas in the corresponding deterministic model it is $c = (K + A)/2$.

Stochastic models reveal that a discount rate below the critical value can substantially reduce the mean time to extinction and

solution²³ of equation (2) with a very high harvesting rate above the threshold and an upper boundary sufficiently large to be effectively inaccessible. This showed that equation (3) is quite accurate, for example within about 1% for values in Fig. 1. (The accuracy of equation (3) is low for N_0 near A , but for discount rates high enough to produce an optimal harvesting threshold in this region the approximation correctly indicates that $c = 0$ is the best strategy.) Differentiating equation (3), using formulae for Y , T and $\bar{N}$ in ref. 21, $\partial \psi / \partial c = 0$ produces the equation for possible values of the harvesting threshold that maximize ψ .

the expected real harvest before extinction (Fig. 2). Even a small discount rate drastically diminishes the maximum expected present value of cumulative harvest before extinction, and little profit can be gained in comparison with immediate harvesting to extinction (Fig. 2); this narrow profit margin will often be eroded by fixed costs of maintaining harvesting operations.

The critical discount rate may also differ between the deterministic and stochastic models. With logistic population dynamics (as in Fig. 2, with $A = 0$), the critical discount rate is identical in the deterministic³ and stochastic models, $\delta^* = r$, the maximum per capita rate of population growth. With an unstable equilibrium at a small population size, the critical discount rate is lower in the stochastic model than in the corresponding deterministic model (Fig. 2).

Thus, even when the discount rate is less than the critical value predicted by deterministic models, the economically optimal strategy will often be immediate harvesting to extinction. These results make a powerful argument that, for the common good²², economic discounting should be avoided in the development of optimal strategies for sustainable use of biological resources.

By enforcing a zero discount rate, if populations can be accurately estimated, threshold harvesting strategies provide a higher expected real harvest before extinction, and a higher mean annual harvest for a longer time, than other strategies²¹. However, threshold strategies entail a high variance in annual harvest, including frequent years of no harvest. Their successful application may therefore require either storage facilities or banking mechanisms to smooth fluctuating harvests. Instituting such harvesting strategies would help to promote a sustainable economy and to maintain biodiversity and ecosystem function on which the continued existence of our own species ultimately depends. □

Received 25 July; accepted 8 September 1994.

1. Clark, C. W. *Science* **181**, 630–634 (1973).
2. Clark, C. W. *Mathematical Bioeconomics* 2nd edn (Wiley, New York, 1990).
3. May, R. M. *Nature* **263**, 91–92 (1976).
4. Allee, W. C., Emerson, A. E., Park, O., Park, T. & Schmidt, K. P. *Principles of Animal Ecology* (Saunders, Philadelphia, 1949).
5. Fowler, C. W. & Baker, J. D. *Rept International Whaling Commission* **41**, 545–554 (1991).
6. Dennis, B. *Natural Resource Modeling* **3**, 481–538 (1989).

7. Ludwig, D., Hilborn, R. & Walters, C. *Science* **260**, 17–36 (1993).
8. Redford, K. H. *BioScience* **42**, 412–422 (1992).
9. Primack, R. B. *Essentials of Conservation Biology* (Sinauer, Sunderland, MA, 1993).
10. Gornbridge, B. (ed.) *Global Biodiversity: Status of the Earth's Living Resources* (Chapman & Hall, New York, 1992).
11. Cumming, D. H. M., Du Toit, R. F. & Stuart, S. N. *African Elephants and Rhinos, Status Survey and Conservation Action Plans* (International Union for the Conservation of Nature and Natural Resources, Gland, Switzerland, 1990).
12. Miller, G. T. Jr *Living in the Environment: An Introduction to Environmental Science* 6th edn (Wadsworth, Belmont, CA, 1990).
13. Le Boeuf, B. in *The Natural History of Año Nuevo* (eds Le Boeuf, B. & Kaza, S.) Ch. 7, 287–325 (Boxwood, Pacific Grove, CA, 1981).
14. Halliday, T. *Vanishing Birds: Their Natural History and Conservation* (Penguin, Harmondsworth, 1980).
15. Van Valen, L. *Evol. Theory* **1**, 1–30 (1973).
16. Jablonski, D. *Science* **231**, 129–133 (1986).
17. Beddington, J. R. & May, R. M. *Science* **197**, 463–465 (1977).
18. May, R. M., Beddington, J. R., Horwood, J. W. & Shepherd, J. G. *Math. biosci.* **42**, 219–252 (1978).
19. Reed, W. J. *Math. Biosci.* **41**, 273–307 (1978).
20. Karlin, S. & Taylor, H. M. *A Second Course in Stochastic Processes* (Academic, New York, 1981).
21. Lande, R., Engen, S. & Saether, B.-E. *Am. Nat.* (in the press).
22. Daley, H. E. & Cobb, J. B. Jr *For the Common Good, Redirecting the Economy Toward Community, the Environment, and a Sustainable Future* (Beacon, Boston, 1989).
23. Keller, H. B. *Numerical Methods for Two-Point Boundary Value Problems* (Blaisdale, New York, 1968).

ACKNOWLEDGEMENTS. We thank W. E. Whitelaw for discussions. This work was supported by grants from the US NSF (R.L.), NINA and the Norwegian Research Council for Science and the Humanities (S.E.), and the Norwegian Directorate for Nature Management (B.-E.S.).

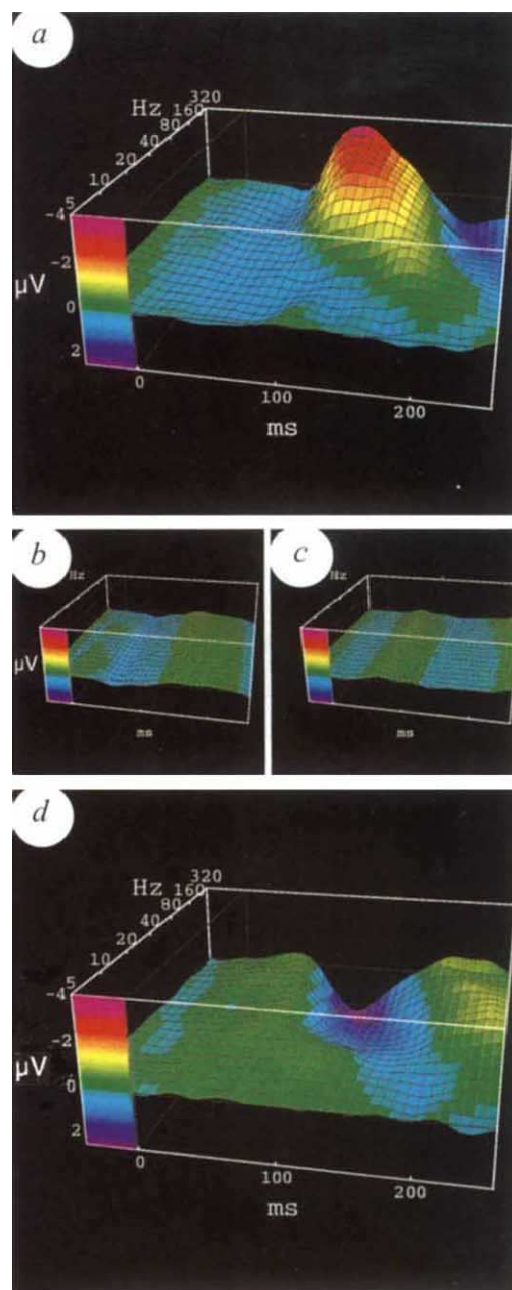
Attentive novelty detection in humans is governed by pre-attentive sensory memory

H. Tiitinen, P. May, K. Reinikainen & R. Näätänen

Cognitive Psychophysiology Research Unit, Department of Psychology, University of Helsinki, SF-00014 Helsinki, Finland

BEING able to detect unusual, possibly dangerous events in the environment is a fundamental ability that helps ensure the survival of biological organisms^{1–3}. Novelty detection requires a memory system that models (builds neural representations of) events in the environment, so that changes are detected because they violate the predictions of the model. The earliest physiologically measurable brain response to novel auditory stimuli is the mismatch negativity⁴, MMN, a component of the event-related potential. It is elicited when a predictable series of unvarying stimuli is unexpectedly followed by a deviating stimulus⁵. As the occurrence of MMN is not usually affected by the direction of attention^{6–7}, MMN reflects the operation of automatic sensory (echoic) memory⁸, the earliest memory system that builds traces of the acoustic environment against which new stimuli can be compared⁵. The dependence of attentive novelty detection on earlier, pre-attentive processes, however, has remained elusive. Previous, related studies^{9–12} seem to suggest a relationship between MMN and attentive processes, although no conclusive evidence has so far been shown. Here we address novelty detection in humans both on a physiological and behavioural level, and show how attentive novelty detection is governed by a pre-attentive sensory memory mechanism.

Ten subjects with normal hearing (mean age, 26 years; 5 males) were presented with 1,000-Hz 'standard' tones randomly interspersed with 'deviant' tones differing in frequency. Novelty, here defined as the magnitude of frequency deviance (the term traditionally used to refer to the property of a stimulus causing attentional, as opposed to arousal, effects^{2,5}), namely the difference in frequency between the standard and deviating stimulus, was varied from 5 to 320 Hz in eight steps. In the first part of the experiment (the event-related potential (ERP) sessions), we measured MMN, with the subject being instructed to read a self-



selected book and to ignore auditory stimuli. In the second part (the novelty detection sessions), we tested the subject's ability to detect the novel stimuli. The subject attended to the same stimuli as those presented in the ERP sessions, and responded to the deviating tones by pressing a response key. Thus the novel stimuli were subject to both pre-attentive and attentive novelty detection.

In Fig. 1, the MMN responses measured from the frontal electrode Fz (Fig. 1a) and the right mastoid (Fig. 1d) are presented. The standard tones elicited only minute, invariant responses (Fig. 1b and c). The peak amplitude of MMN at Fz increased linearly up to 3.5 μ V as a function of the logarithm of the frequency difference (Fig. 3e). MMN was most prominent in the frontal and central scalp areas and it was reliably elicited by frequency differences at 20 Hz (for which one-tailed $t(9) = 2.68$, $P < 0.03$) and above.

The polarity reversal of MMN between frontal and mastoid electrodes suggests that the response originated in the auditory cortex^{13,14}. This was confirmed by magnetoencephalography (MEG) measurements (Fig. 2), from which the sources of activ-

FIG. 1 The results of the ERP sessions. MMN from electrodes Fz (a) and right mastoid (d) for deviance values 5–320 Hz, interpolated between data points along deviance axis by the cubic spline method. The MMN amplitude grew and its latency decreased with increasing deviance. The invariant responses to the standard tones measured from electrodes Fz and right mastoid are depicted in b and c, respectively. METHODS. Measurements were carried out in an acoustically and electrically shielded room. The nose-referenced electroencephalogram (EEG) (sample rate 250 Hz) was averaged and digitally filtered (passband, 2–30 Hz) from 12 channels. One electrooculogram (EOG) channel was used to control eye movements, and epochs with artefacts whose absolute voltage exceeded $75 \mu\text{V}$ at any electrode were omitted. Stimulus blocks (65 dB sound pressure level (SPL), duration 60 ms, including 10 ms rise and fall times) consisting of standard (1,000 Hz, $P=0.95$) and deviant stimuli ($P=0.05$) were presented in a random order through earphones to the subject's left ear. The frequency of the deviant stimulus in a block was either 1,005, 1,010, 1,020, 1,040, 1,080, 1,127, 1,202 or 1,320 Hz. The (onset-to-onset) interstimulus interval was 200 ms. Each subject was presented with 6 blocks of 960 stimuli for each deviance (a total of 437,760 standard and 23,040 deviant tones). The presentation of blocks was counterbalanced. MMN was obtained by subtracting the response to the standard tone from the response to the deviant tone. MMN peak amplitude was measured from grand-average data from Fz as the average voltage in a 20 ms window around the peak, latency as the time of occurrence of the peak, and duration as the time the voltage of the response exceeded $-0.1 \mu\text{V}$.

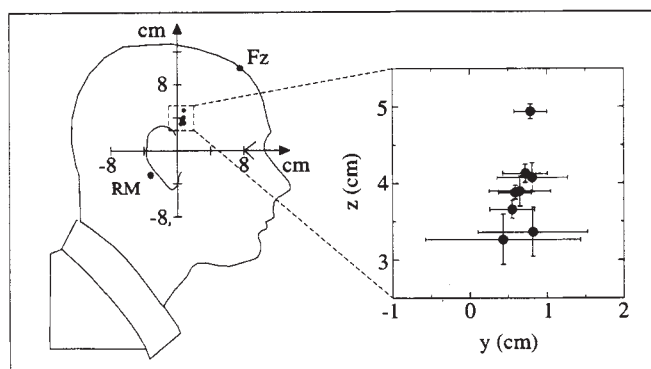


FIG. 2 Localization of the MMN responses with magnetoencephalography (MEG) for a representative subject. The activated brain area, estimated with equivalent current dipoles (ECDs, denoted by black dots, with bars indicating 95% confidence limits for 100 stimulus presentations per deviance), was in the approximate location of supratemporal primary auditory cortex. Although it was highly localized, MMN source location showed no systematicity as a function of deviance (5 to 10 Hz deviations resulted in ECDs with largest confidence limits). The co-ordinate system origin is set at the preauricular point. The measurement locations Fz, RM (right mastoid) of the main experiment are also shown. Three subjects of the main experiment were studied, with no statistical differences in response locations.

METHODS. Magnetic fields were recorded with a helmet-shaped whole-head Neuromag-122 SQUID magnetometer²⁰ (passband 0.03–90 Hz, sample rate 297 Hz, stimulus delivery through a plastic tube and an ear piece). The experimental arrangement is described in Fig. 1 legend.

ity for the eight deviances were found to be highly localized in an area ~ 2 cm below the scalp and 3 cm above the preauricular point. Also, it is unlikely that the responses were contaminated by other ERP components such as N1 or N2b, because N1 has been shown to overlap MMN only when deviance values exceed 50 per cent¹⁵ and N2b is observed only when subjects attend to stimuli, and its polarity does not reverse at the mastoids¹⁶.

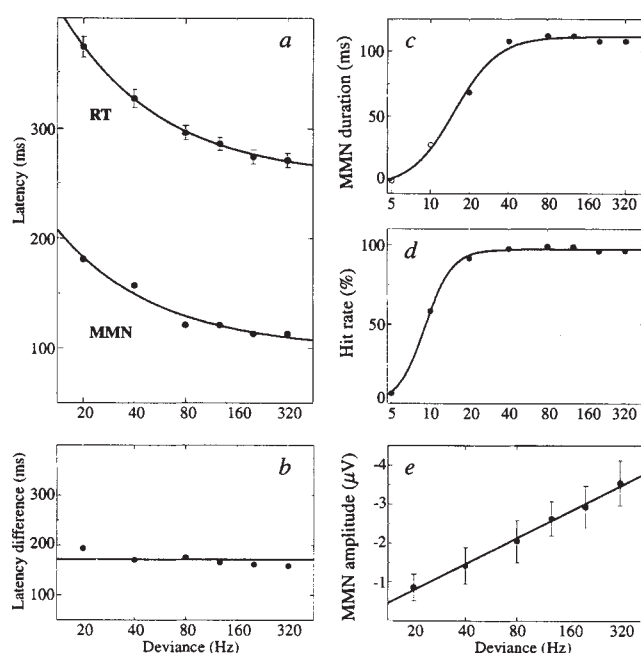


FIG. 3 Results of the main experiment. a, Both the MMN peak latency from Fz, and behavioural reaction time (RT) decreased monotonically as a function of change in tone frequency (data have been fitted with an exponential decay function; bars indicate standard error of the mean). b, MMN peak latency has been subtracted from behavioural RT, demonstrating that the difference was constant. With increasing deviance both the MMN duration from Fz (c) and the behaviourally obtained hit rate (d) increased until reaching a plateau (data fitted with a sigmoidal function). Durations of responses to 5- and 10-Hz deviances are denoted by open circles, with the former arbitrarily defined as zero. e, MMN peak amplitude increased linearly as a function of the logarithm of frequency deviance (bars indicate standard error of the mean). METHODS. The subject was, on a separate day, presented with two stimulus blocks per deviance (about 75 presentations per deviance) and asked to respond accurately and rapidly to deviant tones by pressing a response key. The time window for an acceptable response was defined as 200–700 ms after stimulus onset, and the hit rate was defined as the percentage of target tones found by the subject. This procedure took into account 95% of all the responses.

Figure 3a shows that the peak latency of MMN decreased monotonically (from 180 to 110 ms) and that the subject's reaction time in the novelty detection sessions also decreased monotonically (from 370 to 270 ms) as a function of frequency difference. Reaction time changed at the same rate as MMN latency, the two measures being highly correlated (Pearson $r=0.986$, two-tailed $P<0.001$) and the difference between these two being constant (Fig. 3b).

MMN duration increased and reached a plateau of 100 ms at ~ 40 Hz frequency difference (Fig. 3c). The subject's ability to detect novel stimuli, as reflected in the hit rate, also improved until it reached a plateau at ~ 40 Hz frequency difference (Fig. 3d). Thus, the hit rate correlated with the MMN duration function in asymptotic behaviour (Pearson $r=0.934$, two-tailed $P<0.01$).

In conclusion, our main finding was that the latency of MMN predicted the behavioural response latency extremely well. Thus, attentive behavioural response latency changes can be explained as originating from the pre-attentive sensory memory mechanism. Changes, due to task difficulty, in sensory-memory processing time seem then to be linearly transferred to subsequent (perceptual, attentive, motor) processes of constant duration. Whereas changes in reaction time as a function of task difficulty are well established¹⁷, our results show systematically that reaction time changes for the oddball paradigm are governed

solely by early sensory memory processes, which must therefore be fundamental in determining the appropriate behaviour of the organism.

The duration function (rarely addressed in ERP studies) of MMN resembles the hit-rate function in that it rapidly reached a plateau. This suggests that what determines whether a novel stimulus is consciously detected might depend on the temporal length of the neural signal generated by the sensory memory mechanism; that is, stimulus deviance could be transformed into a time code which is then broadcast to attentional mechanisms. However in conditions of low discrimination, MMN duration reaches a plateau slightly later, implying that some other mechanism is enhancing the hit rate.

MMN is reliably elicited by deviances at and above 20 Hz and its amplitude is directly proportional to the logarithm of frequency difference. As this relationship between frequency difference and response amplitude parallels the previously observed behavioural relationship between sound frequency difference and perceived pitch difference^{18,19}, the perception of pitch differences might therefore be reflected in the physiological measure of MMN amplitude. □

Received 28 July; accepted 20 September 1994.

- James, W. *The Principles of Psychology* Vol. 1, 402–458 (Dover, New York, 1950).
- Sokolov, E. N. & Vinogradova, O. S. *Neuronal Mechanisms of the Orienting Reflex* 217–235 (Erlbaum, New York, 1975).
- Kohonen, T. *Self-Organization and Associative Memory* 241–248 (Springer, Amsterdam, 1989).
- Näätänen, R., Gaillard, A. W. K. & Mäntysalo, S. *Acta psychologica* **42**, 313–329 (1978).
- Näätänen, R. *Attention and Brain Function* 136–199 (Erlbaum, New Jersey, 1992).
- Woldorff, M., Hackley, S. A. & Hillyard, S. A. *Psychophysiology* **28**, 30–42 (1991).
- Näätänen, R. *Psychophysiology* **28**, 478–484 (1991).
- Neisser, U. *Cognitive Psychology* 199–218 (Appleton Century Crofts, New York, 1967).
- Sams, M. et al. *Electroenceph. clin. Neurophys.* **62**, 437–448 (1985).
- Winkler, I., Reinikainen, K. & Näätänen, R. *Percept. Psychophys.* **53**, 443–449 (1993).
- Novak, G., Ritter, W., Vaughan, H. G. Jr & Witznitzer, M. L. *Electroenceph. clin. Neurophys.* **75**, 255–275 (1990).
- Novak, G., Ritter, W. & Vaughan, H. G. Jr *Psychophysiology* **29**, 398–411 (1992).
- Hari, R. et al. *Neurosci. Lett.* **50**, 127–132 (1984).
- Tiitinen, H. et al. *Psychophysiology* **30**, 537–540 (1993).
- Scherg, M., Vajsaar, J. & Picton, T. W. J. *cogn. Neurosci.* **2**, 336–355 (1989).
- Ritter, W. et al. *Electroenceph. clin. Neurophys.* **83**, 306–322 (1992).
- Keele, S. in *Handbook of Perception and Human Performance* Vol. 2 (eds Boff, K. R., Kaufman, L. & Thomas, J. P.) 30.1–30.60 (Wiley, New York, 1986).
- Wier, C. C., Jesteadt, W. & Green, D. M. J. *acoust. Soc. Am.* **61**, 178–184 (1977).
- Stevens, S. S. & Davis, H. *Hearing* 69–109 (Wiley, New York, 1966).
- Ahonen, A. I. et al. in *Proc. Satell. Symp. Neurosci. Technol. 14th Int. Conf. IEEE Engng Med. Biol. Soc.* (eds Dittmar, A. & Froment, J. C.) 16–20 (Lyon, 1992).

ACKNOWLEDGEMENTS. This work was supported by The Academy of Finland, The Finnish Cultural Foundation and The Emil Aaltonen Foundation. We thank M. Huotilainen and J. Sinkkonen for their help in preparing the manuscript, and the Low Temperature Laboratory of the Helsinki University of Technology and Neuromag Ltd for the use of the magnetometer.

Amyloid-associated proteins α_1 -antichymotrypsin and apolipoprotein E promote assembly of Alzheimer β -protein into filaments

Jianyi Ma, Ann Yee, H. Bryan Brewer Jr*,
Saumya Das & Huntington Potter†

Department of Neurobiology, Harvard Medical School,
220 Longwood Avenue, Boston, Massachusetts 02115, USA

* National Heart and Blood Institute, National Institutes of Health,
Bethesda, Maryland 20892, USA

THE protease inhibitor α_1 -antichymotrypsin and the lipid transport protein apolipoprotein E (apoE) are intimately associated with the 42-amino-acid β -peptide ($A\beta$) in the filamentous amyloid deposits of Alzheimer's disease^{1–3}. We report here that these two amyloid-associated proteins serve a strong stimulatory role in the polymerization of $A\beta$ into amyloid filaments. Addition of either α_1 -antichymotrypsin or apoE to the $A\beta$ peptide promoted a 10- to 20-fold increase in filament formation, with apoE-4, the isoform recently linked to the development of late-onset Alzheimer's disease, showing the highest catalytic activity. These and other experiments suggest that Alzheimer amyloid deposits arise when $A\beta$ is induced to form filaments by amyloid-promoting factors (pathological chaperones) expressed in certain brain regions.

A number of lines of investigation have indicated that α_1 -antichymotrypsin (ACT) has a specific affinity for the $A\beta$ peptide^{1,4–9} and there are comparable data for apoE^{2,3,10,11}. These results suggested that one or both of these amyloid-associated proteins might play a direct role in the aetiology of Alzheimer's disease by functioning as a 'pathological chaperone' to promote the formation of amyloid filaments^{3,7}. Monomers of the predominant amyloid protein found in neuritic plaques and vascular amyloid, $A\beta_{1-42}$ (refs 12, 13), assemble spontaneously into high-molecular-mass amyloid-like structures, as monitored by light scattering and by electron microscopy (Fig. 1 and refs 14, 15). Over the course of several days, filaments ~7 nm wide

and up to many hundreds of nanometres long steadily increased in number.

Although $A\beta$ can assemble into amyloid-like filaments spontaneously, the data shown in Figs 2 and 4 show that filaments form much more rapidly in the presence of the amyloid-associated protein ACT. Instead of days, filaments formed in a matter of hours and grew to very great lengths, frequently traversing an entire electron microscope grid square (100 $\mu\text{m} \times 100 \mu\text{m}$). Quantitative analysis done by counting random filament cross-overs (a measure of both filament number and length) indicated that ACT accelerated the rate of filament formation at least 10-fold (Figs 2 and 4).

During the course of these studies, it became apparent that apoE, like ACT, is a potential amyloid-promoting factor, or pathological chaperone³. As in the case of ACT, apoE is both

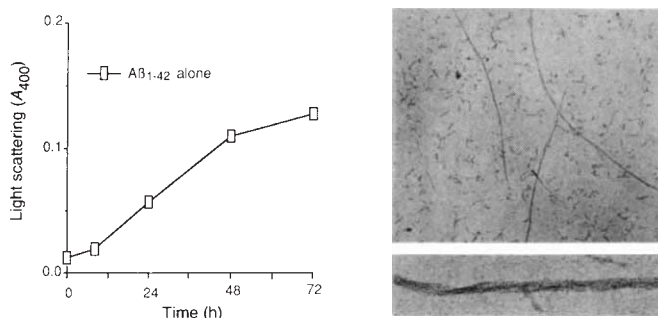


FIG. 1 The Alzheimer $A\beta$ peptide forms amyloid filaments *in vitro*. Biosynthetic $A\beta_{1-42}$ was incubated at a concentration of 80 μM in 100 μl of 10 μM Tris-HCl, pH 7.0, at 22 °C for the indicated times. Assembly of $A\beta$ into high- M_r filamentous structures was monitored by light scattering at 400 nm (left), which measures the total mass of material in macromolecular form. At 18 h, filaments were applied to carbon-coated Formvar on 200 mesh copper grids, washed, dried, negatively stained with 3% uranyl acetate and visualized in a JEOL-100CX electron microscope. The filaments shown in the electron micrograph (right) are representative of the few products of this polymerization reaction; magnification: top right, $\times 18,000$; bottom right, $\times 84,000$. Because $A\beta_{1-42}$ is prone to self-aggregation and because filtration fails to remove small filaments, aliquots of each lyophilized preparation of $A\beta_{1-42}$ were tested by Hoffman-LaRoche and by us for the presence of amyloid filaments. Only preparations devoid of filaments by electron microscopy were used for experiments.

† To whom correspondence should be addressed.

present in Alzheimer amyloid deposits and binds tightly to the A β protein *in vitro*^{2,3,10,11}. Moreover, epidemiology has indicated that the development of late-onset Alzheimer's disease depends strongly on, and may be caused by, a particular apoE allele (apoE-4) carried by affected individuals^{11,16}. We therefore examined the relative ability of purified human apoE-2, apoE-3 and apoE-4 (ref. 17) to promote A β filament formation. Like ACT, apoE strongly promoted A β polymerization, with the apoE-4 isoform being the most, and apoE-2 the least, active (Figs 3 and 4). Interestingly, apoE-2, when combined with apoE-4, inhibited the formation of amyloid filaments, consistent with its apparent protective role *in vivo*¹⁸. Two other apolipoproteins, AI and AII, were inactive as amyloid-promoting factors (Fig. 4).

Both ACT and apoE interact directly with the A β peptide^{2,3,10,11}. Amino acids 1–12 of A β closely resemble the

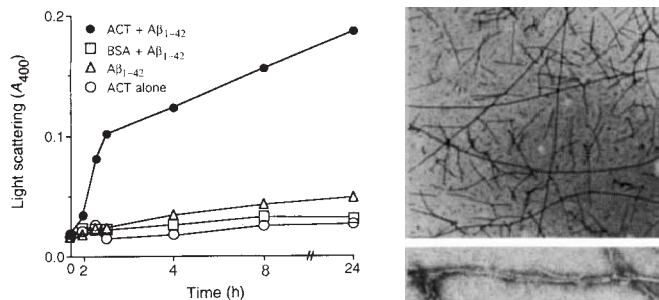


FIG. 2 The Alzheimer amyloid-associated protein ACT promotes the assembly of A β protein into filaments. The A β _{1–42} peptide was incubated as described in the legend to Fig. 1 in the presence or absence of α_1 -antichymotrypsin (ACT; CalBiochem) or a control protein (BSA; USB) at a molar ratio of 200:1. Total mass of material in macromolecular form was monitored by light scattering (left). Electron microscopy revealed large numbers of long, ~7-nm-wide, amyloid-like filaments, sometimes having a double-stranded appearance, and many shorter filaments (top right, magnification $\times 17,400$). ACT promoted an increase in A β filament formation of at least 10-fold, but contained and produced no filaments itself (see also Fig. 4). High-magnification electron microscopy (lower right, $\times 81,200$) revealed filaments similar to those formed by A β alone and shown in Fig. 1 and those found in tissue sections from Alzheimer's disease brain.

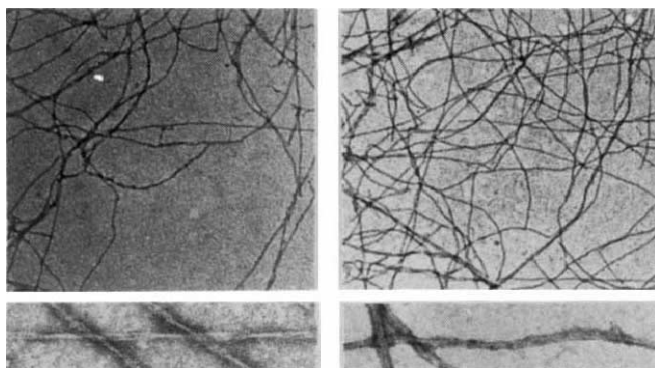


FIG. 3 The amyloid-associated protein apoE promotes the assembly of A β peptide into filaments. The A β _{1–42} peptide was incubated in the presence of purified human apolipoproteins E-3 and E-4 at a molar ratio of 200:1, as described in the legends to Figs 1 and 2. ApoE-3 and E-4 proteins were prepared as described in ref. 17. Electron microscopic examination of the reaction products of A β with apoE-3 (left) or apoE-4 (right) indicated that apoE-3 was the more active in promoting formation of filaments that were indistinguishable in appearance from those generated by A β alone or in the presence of ACT (compare with Figs 1 and 2). Magnification: left top, $\times 21,000$, bottom, $\times 98,000$; right top, $\times 21,000$, bottom, $\times 98,000$.

conserved active site of serine proteases, the natural 'target' for protease inhibitors such as ACT, and *in vitro* binding experiments have confirmed that this region is specifically recognized and bound by ACT, apparently as a pseudosubstrate^{7–9}.

In contrast to ACT, apoE is found associated with all cerebral and systemic amyloids, apparently reflecting a more general affinity for the β -pleated-sheet conformation characteristic of amyloid filaments³. Indeed, apoE interacts primarily with amino acids 12–28 of A β , a sequence predicted to form a β -pleated sheet¹¹. Thus the two amyloid-associated proteins, ACT and apoE, each recognize and bind to A β to promote filament formation but at different sites and with different degrees of specificity.

It has become increasingly clear that the 42-amino-acid form of the A β peptide predominates in the amyloid deposits of Alzheimer's disease^{12,13}. We therefore chose to do experiments primarily with A β _{1–42}, and have found that ACT and apoE promote filament formation. Our earlier experiments had indicated that ACT and apoE were less effective at promoting polymerization of A β _{1–40}. Indeed, under conditions of high concentrations of peptide and ACT, Fraser *et al.*⁹ even found a negative effect of ACT on A β _{1–40} fibril formation, possibly reflecting a greater affinity of A β _{1–40} for ACT than for itself. The importance of studying A β _{1–42} is underscored by the finding that familial Alzheimer's disease mutations in the APP gene increase the production of A β _{1–42} in APP-transfected cells compared to A β _{1–40} (refs 19, 20).

One unexplained feature of Alzheimer's disease neuropathology is its regional specificity. The A β peptide is produced in most cell types, and its aggregation to form amorphous deposits occurs in many brain regions, but mature amyloid and neuronal cell death exhibits a much more restricted localization^{19–24}. Our data suggest that the region-specific deposition of filamentous amyloid may reflect the regional availability not of the A β peptide itself, but of pathological chaperones such as ACT and

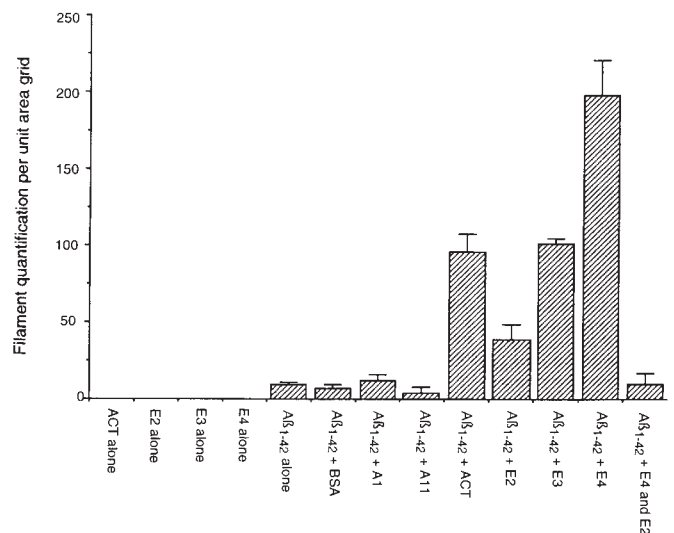


FIG. 4 Quantitative electron microscopic analysis of A β filament formation promoted by pathological chaperones. Four randomly chosen electron micrographs of each reaction product after 18 h of incubation as in Figs 1–3 were analysed by counting the total number of filament crossovers per unit area of the grid. The results $\pm$ s.e.m. of one experiment are shown: apoE-4 was most active in promoting filament formation; ACT and apoE-3 exhibited strong intermediate activity; apoE-2 had lower activity and inhibited filament formation by apoE-4. Three independent experiments gave similar results. The control proteins BSA and apolipoproteins AI and AII were inactive as amyloid-promoting agents. No filaments formed when any of the tested proteins were incubated in the absence of A β peptide.

apoE, which promote the formation of the amyloid filaments. For instance, ACT messenger RNA and protein are greatly increased in those areas of Alzheimer brain prone to filamentous amyloid deposition, being present in the circulation and synthesized in reactive astrocytes in response to interleukin-1 released from microglia^{1,8,25,26,29}. Similarly, increased numbers of plaques are found in individuals with Alzheimer's disease expressing the rare apoE-4 isoform^{27,28}. Together these results indicate that a multi-step pathogenic pathway leads to Alzheimer neuropathology. Rate-limiting steps in this pathway, such as the interaction between ACT and/or apoE and A β , might serve as targets for therapeutic intervention.

Note added in proof: Since this Letter was submitted, the influence of apoE on A β ₁₋₄₀ filament formation has been independently confirmed by T. Wisniewski *et al.*³⁰ and by D. A. Sanan *et al.*³¹. □

Received 5 May; accepted 14 September 1994.

1. Abraham, C. R., Selkoe, D. J. & Potter, H. *Cell* **52**, 487–501 (1988).
2. Namba, Y., Tomonaga, M., Kawasaki, H., Otomo, E. & Ikeda, K. *Brain Res.* **541**, 163–166 (1991).
3. Wisniewski, T. & Frangione, B. *Neurosci. Lett.* **135**, 235–238 (1992).
4. Abraham, C. R., Shirahama, T. & Potter, H. *Neurobiol. Aging* **11**, 123–129 (1990).
5. Picken, M. M. *et al. J. Neuropathol. exp. Neurol.* **49**, 41–48 (1990).
6. Rozemuller, J. M. *et al. Acta Neuropathol.* **82**, 200–207 (1991).

7. Potter, H., Abraham, C. R. & Dressler, D. in *Alzheimer's Disease: Basic Mechanisms, Diagnosis and Therapeutic Strategies* (eds Iqbal, K. *et al.*) 275–279 (Wiley, New York, 1991).
8. Potter, H. *et al. Ann. N.Y. Acad. Sci.* **674**, 161–173 (1992).
9. Fraser, P. E. *et al. J. Neurochem.* **61**, 298–305 (1993).
10. Wisniewski, T., Golabek, A., Matsubara, E., Ghiso, J. & Frangione, B. *Biochem. biophys. Res. Commun.* **192**, 359–365 (1993).
11. Strittmatter, W. J. *et al. Proc. natn. Acad. Sci. U.S.A.* **90**, 1977–1981 (1993).
12. Roher, A. E. *Proc. natn. Acad. Sci. U.S.A.* **90**, 10836–10840 (1993).
13. Iwatsubo, T. *et al. Neuron* **13**, 45–53 (1994).
14. Castaño, E. M. *et al. Biochem. biophys. Res. Commun.* **141**, 782–789 (1986).
15. Kirschner, D. A. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 6953–6957 (1987).
16. Corder, E. H. *et al. Science* **261**, 921–923 (1993).
17. Gregg, R. E. *et al. J. clin. Invest.* **78**, 815–821 (1986).
18. Corder, E. H. *et al. Nature Genet.* **7**, 180–184 (1994).
19. Suzuki, N. *et al. Science* **264**, 1336–1340 (1994).
20. Dovey, H. F., Suomensaari-Chrysler, S., Lieberburg, I., Sinha, S. & Keim, P. S. *NeuroReport* **4**, 1039–1042 (1993).
21. Haass, C. *et al. Nature* **359**, 322–325 (1992).
22. Probst, A., Brunnenschweiler, H., Lautenschlager, C. & Ulrich, J. *Acta Neuropathol.* **74**, 133–141 (1987).
23. Seubert, P. *et al. Nature* **359**, 325–327 (1992).
24. Shoji, M. *et al. Science* **258**, 126–129 (1992).
25. Pasternack, J. M., Abraham, C. R., Van Dyke, B., Potter, H. & Younkin, S. G. *Am. J. Pathol.* **135**, 827–834 (1989).
26. Koo, E. H., Abraham, C. R., Potter, H., Cork, L. C. & Price, D. L. *Neurobiol. Aging* **12**, 495–501 (1992).
27. Rebeck, G. W., Reiter, J. S., Strickland, D. K. & Hyman, B. T. *Neuron* **11**, 575–580 (1993).
28. Schmechel, D. E. *et al. Proc. natn. Acad. Sci. U.S.A.* **90**, 9649–9653 (1993).
29. Das, S. & Potter, H. *Neuron* (in the press).
30. Wisniewski, T. *et al. Am. J. Path.* (in the press).
31. Sanan, D. A. *et al. J. clin. Invest.* **94**, 860–869 (1994).

ACKNOWLEDGEMENTS. We thank G. Huber, H. Döbeli, N. Draeger, M. Zulauf and B. Seilheimer of Hoffman-LeRoche for β ₁₋₄₂ peptide and for advice and discussion and D. Dressler for excellent suggestions on the manuscript. The work was supported by grants and fellowships from the NIH, the Freudenberg family and the Howard Hughes Medical Institute.

Pinopsin is a chicken pineal photoreceptive molecule

Toshiyuki Okano, Tôru Yoshizawa*
& Yoshitaka Fukada†

Department of Pure and Applied Sciences,
College of Arts and Sciences, The University of Tokyo, Komaba 3-8-1,
Meguro-Ku, Tokyo 153, Japan

* Department of Information Systems Engineering,
Faculty of Engineering, Osaka Sangyo University, Nakagaito 3-3-1,
Daito-Shi, Osaka 574, Japan

In avian pinealocytes, an environmental light signal resets the phase of the endogenous circadian pacemaker that controls the rhythmic production of melatonin¹⁻⁶. Investigation of the pineal phototransduction pathway should therefore reveal the molecular mechanism of the biological clock. The presence of rhodopsin-like photoreceptive pigment^{4,5,7-9}, transducin-like immunoreaction¹⁰, and cyclic GMP-dependent cation-channel activity¹¹ in the avian pinealocytes suggests that there is a similarity between retinal rod cells and pinealocytes in the phototransduction pathway. We have now cloned chicken pineal cDNA encoding the photoreceptive molecule, which is 43–48% identical in amino-acid sequence to vertebrate retinal opsins. Pineal opsin, produced by transfection of complementary DNA into cultured cells, was reconstituted with 11-*cis*-retinal, resulting in formation of a blue-sensitive pigment ($\lambda_{\max} \approx 470$ nm). In the light of this functional evidence and because the gene is specifically expressed only in the pineal gland, we conclude that it is a pineal photosensor and name it pinopsin.

For the initial screening of a λ gt11 chick pineal cDNA library, we used cDNA probes encoding chicken retinal photoreceptive pigments, the red-¹² and violet-sensitive cone pigments¹³. One of the isolated cDNA clones was identical with the 3' region of chicken 'red' cDNA¹², suggesting that a red-sensitive pigment was present in the pinealocytes. We obtained two partial cDNA clones encoding amino-acid sequences homologous but not identical to the carboxy- and amino-terminal regions, respec-

tively, of known visual pigments. A pair of oligonucleotide primers possibly corresponding to 5' and 3' flanking regions of the pineal protein was synthesized for polymerase chain reaction (PCR) with a chick pineal cDNA template. The reaction amplified a full-length PCR clone (V2L) encoding a photoreceptive pigment named 'pinopsin' (pineal opsin) composed of 351 amino acids (Fig. 1). Using V2L as a probe, a full-length cDNA clone (Xd) was also isolated from the pineal cDNA library.

Hydropathy analysis of the sequence predicts a heptahelical structure of pinopsin, being a member of the G-protein-coupled receptor family. Of several key conserved residues, a lysine (K288) that can attach to a retinal chromophore and lying in the middle of the seventh helix (Fig. 1) is a common feature of all photoreceptive molecules cloned from the retina so far. Northern blot analysis indicates that pinopsin messenger RNA is transcribed only in the pineal gland (Fig. 2a), and that it is more abundant than the red-sensitive pigment (Fig. 2b). No signal for the expression of pinopsin was detected in retinal poly(A)⁺ RNA up to 10 μ g (not shown). To our knowledge, pinopsin is the first example of an extraretinal photoreceptive molecule.

To reveal the evolutionary background of pinopsin, a phylogenetic tree of vertebrate photoreceptive pigments was constructed on the basis of amino-acid identity¹⁴ (Fig. 3). Vertebrate retinal pigments can be classified into four groups, L, S, M1, and M2 (refs 13, 15); rhodopsins (group Rh) diverged from group M2 later. Every pair of sequences clustered in the same group has more than 70% amino-acid identity. It should be noted, however, that pinopsin is 43–48% identical in sequence to vertebrate retinal pigments, and thus forms a novel group termed 'P' (Fig. 3). The tree shows that pinopsin has diverged from an ancestor of cone pigments (group L) before the divergence of rhodopsins in group M2, suggesting strongly that pinopsin is not a rod-type but a cone-type pigment. Among vertebrate photoreceptive pigments, pinopsin has a unique sequence in the C-terminal region (residues 320–351; Fig. 1) which contains only a few potential phosphorylation sites (serine/threonine residues). In addition, pinopsin, like invertebrate pigments, lacks two amino acids between Ser 190 and Asn 191 as compared with vertebrate visual pigments. These facts imply a possible divergence of pinopsin from a common ancestor of vertebrate visual pigments rather than from an ancestor of group L.

† To whom correspondence should be addressed.

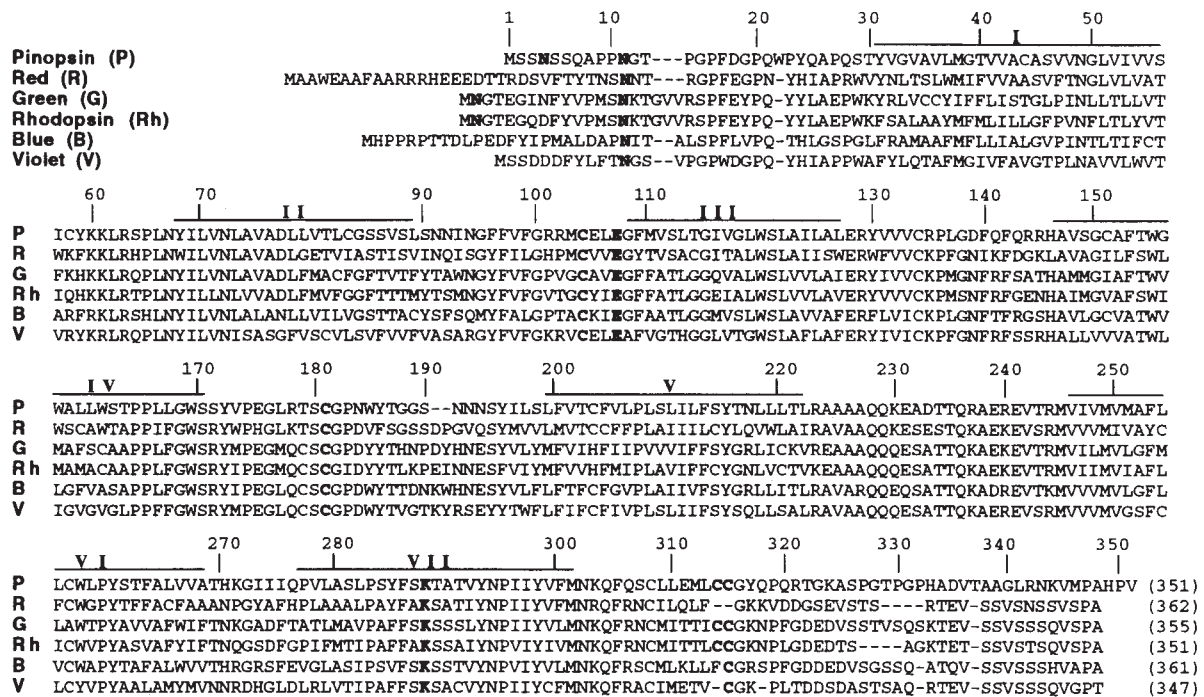
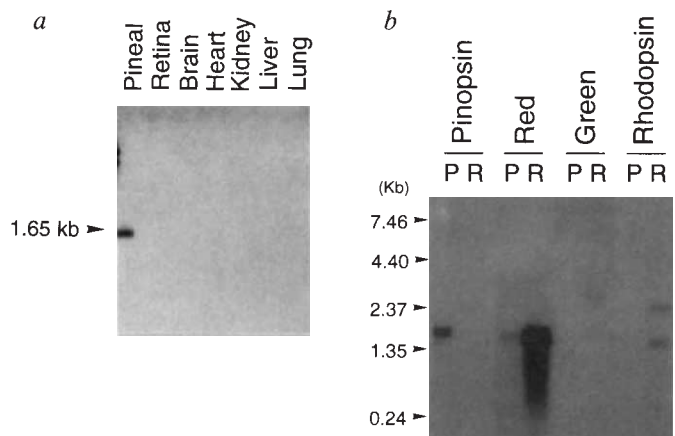


FIG. 1 The deduced amino-acid sequence of pinopsin. Gaps (dashes) were inserted in the sequences for the optimal alignment of the sequences of pinopsin and chicken visual pigments^{12,13,19}. Putative transmembrane domains I–VII are indicated by horizontal lines, and bold-face characters represent potential sites for N-glycosylation (Asn 4 and Asn 11 in pinopsin), intramolecular disulphide bond (Cys 104 and Cys 181), retinylidene Schiff's base counterion (Glu 107), Schiff's base linkage with 11-cis-retinal (Lys 288), and palmitoylation (Cys 314 and Cys 315).

METHODS. Total RNA extracted from 21 one-day-old chick pineal glands was run on an oligo(dT) cellulose column (Collaborative Research Inc.) to separate poly(A)⁺ RNA²⁰. A chick pineal cDNA library was constructed in λ gt11 phage using this poly(A)⁺ RNA and a cDNA cloning kit (Amersham). In the initial screening, plaques of 9.6×10^4 recombinants were transferred to nylon membranes (Hybond-N⁺, Amersham) for hybridization with a mixture of ³²P-random-labelled cDNAs (Random primer labelling kit, Takara) encoding chicken violet- and red-sensitive cone visual pigments^{12,13}. Of the cDNA clones isolated, P3a encoded residues Tyr 281–Ala 362, corresponding to the C-terminal region of chicken red-

sensitive pigment; P4b and P3b encoded sequences homologous to the N and C termini of retinal photoreceptive pigments, respectively. A pair of primers, pampN1 (5'-GCTTTATAAGAGGAGGCTTTGGGAG-3': part of P4b) and pampR1 (5'-GGCTTCATAGCAAGCTC-3': part of P3b), which correspond to 5'- and 3'-flanking regions respectively, were synthesized and used for PCR amplification of the entire coding region. PCR employed Vent polymerase (New England Biolabs), template chick pineal cDNAs, and 35-cycle reactions of 1 min at 94 °C, 1 min at 56 °C, and 1 min at 72 °C, with the final extension reaction lasting for 10 min. The reaction amplified a single DNA fragment, which was then cloned into pBluescript II KS(+) plasmid vector (Stratagene) for sequence analysis by dideoxynucleotide chain termination (Sequenase Ver.2.0, United States Biochem.). We sequenced both strands of six independent clones and confirmed no PCR error. The chick pineal cDNA library constructed in λ gt11 phage (plaques of 4×10^5 recombinants) was screened by ³²P-random-labelled PCR clone V2L (1,484 bp), resulting in isolation of a cDNA clone Xd (1,582 bp) containing the entire coding region. The nucleotide sequence of Xd matched that of V2L exactly.

FIG. 2 a, Northern blot analysis of pinopsin mRNA in various chick tissues. Total RNA was extracted from each tissue²⁰, and subjected to oligo(dT) cellulose column (Collaborative Research). The poly(A)⁺ RNA (2 μ g) was incubated for 1 h at 50 °C in 10 mM sodium phosphate buffer (pH 7.0) containing 1 M glyoxal and 50% (w/v) dimethylsulphoxide, electrophoresed on a 1% agarose gel, transferred onto a nylon membrane and hybridized with ³²P-random-labelled clone V2L (1×10^6 d.p.m.) for 20 h at 42 °C in 10 mM sodium phosphate (pH 6.5) containing 5 $\times$ SSC, 0.25% skim milk, 47% formamide, 100 μ g ml⁻¹ denatured salmon sperm DNA and 10% dextran sulphate. Finally, the membrane was washed at 50 °C with 0.1 $\times$ SSC and 0.1% SDS, and autoradiographed. The size of the hybridized band (1.65 kb), estimated by comparison with a 0.24–9.5 kb RNA ladder (BRL), agreed well with that of the full-length cDNA clone Xd (1,582 bp). b, Northern blot analysis of mRNAs of pinopsin and retinal pigments in chicken pineal gland and retina. The poly(A)⁺ RNA (4 μ g for each lane) isolated from the pineal gland (P) or retina (R) was glyoxalated, electrophoresed, and transferred to a nylon membrane as before. Membranes were hybridized with ³²P-random-labelled clones (1×10^7 d.p.m.) encoding chicken pinopsin, red-sensitive pigment¹², green-sensitive pigment¹³ and rhodopsin¹⁹, then washed at 66 °C with 0.1 $\times$ SSC and 0.1% SDS and autoradiographed.



To investigate the spectral sensitivity as a functional assay, pinopsin was produced in cultured cells. The full-length cDNA clone V2L was subcloned into a eukaryotic expression vector pREP4 and transfected into 293EBNA cells, which were then cultured for 65 h. The membrane fraction of the transfected cells was supplemented with an excess of 11-*cis*-retinal, the most likely chromophore of chicken pineal photoreceptive molecule(s)^{9,16}. Pinopsin thus regenerated in the membranes was extracted with a buffer containing CHAPS and egg yolk phosphatidylcholine, an effective solubilizer that preserves the spectral sensitivities of unstable photoreceptive molecules¹⁷. The difference absorption spectrum between the regenerated pinopsin and its photo-

products, apoprotein plus retinal-oxime, showed a $\lambda_{\max} \sim 470$ nm (Fig. 4). Pinopsin is a blue-sensitive photoreceptive pigment in chicken pinealocytes.

Chicken pinealocytes have two distinct but closely related phototransduction pathways^{6,18}. One contributes to the acute inhibition by light of the serotonin *N*-acetyltransferase enzyme that is responsible for the melatonin rhythm of chicken pinealocytes. The other triggers a light-induced phase-shift of the circadian pacemaker. Green light (500–520 nm) is most effective on the former pathway⁴. The reported action spectrum is red-shifted and broader in shape than the absorption (difference) spectrum of pinopsin. We can speculate that, in addition to pinopsin, the

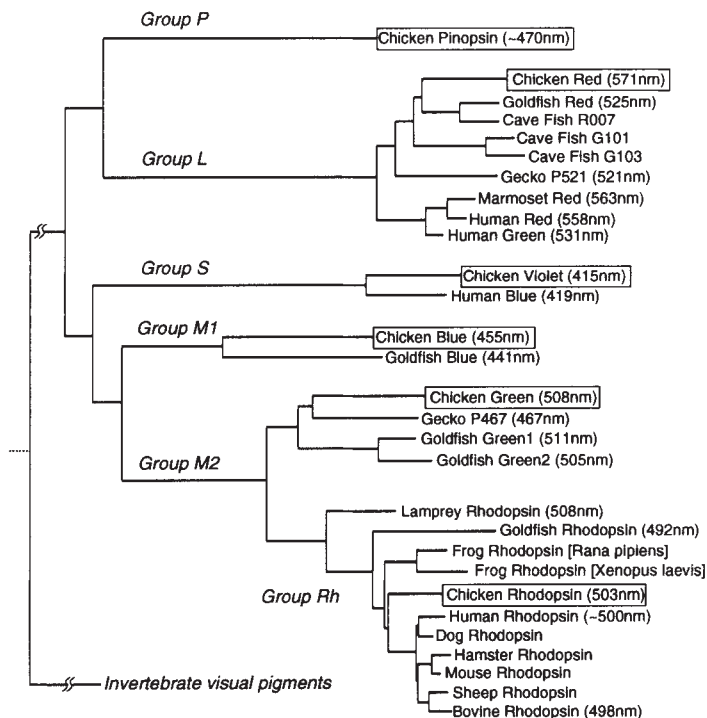
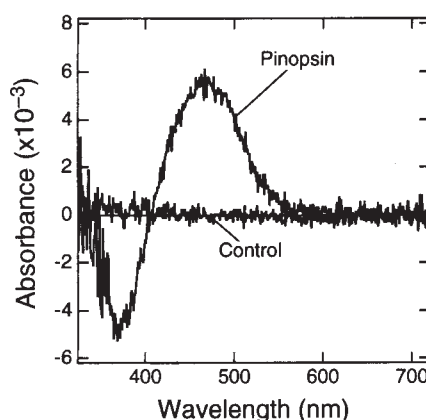


FIG. 3 Phylogenetic tree of vertebrate photoreceptive pigments constructed on the basis of amino-acid identities. The tree was constructed by the neighbour-joining method¹⁴. The deepest root of the tree was determined by constructing another expanded tree which included several invertebrate opsins as outgroup. N- and C-terminal regions (corresponding to residues 1–29 and 307–351 in pinopsin) and the loop region (residues 223–237 in pinopsin) between fifth and sixth transmembrane segments were eliminated from the calculation because the sequence alignment required insertion of many gaps in these regions of invertebrate opsins. The evolutionary distance ($k = -\ln(1 - K)$) was estimated by calculating the amino-acid difference ($K = 1 - (\% \text{ identity})/100$). Lengths of horizontal lines represent branch lengths calculated from the evolutionary distances (k). Amino-acid sequences were obtained from GenBank. Photoreceptive molecules are followed by their absorption maxima in parentheses; those from chicken are boxed. Groups L, M1, M2 and S include long-, middle-1-, middle-2- and short-wavelength-sensitive pigments, respectively¹³.

FIG. 4 Difference absorption spectrum of pinopsin expressed in 293EBNA cells. Pinopsin expressed in pREP4-V2L-plasmid-transfected cells was regenerated with exogenously added 11-*cis*-retinal and then extracted with 0.75% CHAPS. The extract (0.2 ml) was mixed with 2 μ l 1M hydroxylamine-HCl neutralized with NaOH, followed by incubation for 30 min to convert free retinal and random Schiff's base into retinal oxime completely¹⁷. The difference absorption spectrum (pinopsin) was measured at 4 °C before and after irradiation of the sample with orange light (>520 nm) for 5 min. As a control, 293EBNA cells were transfected with pREP4-CAT plasmid (Invitrogen) instead of pREP4-V2L.

METHODS. A cDNA fragment V2L containing the entire coding region of pinopsin was inserted into the expression vector pREP4 (Invitrogen) to construct pREP4-V2L. 400 μ g pREP4-V2L were transfected into 293EBNA cells (Invitrogen) cultured in 20 plates (10 cm diameter) by the calcium phosphate method²¹. Cells were grown for 65 h at 37 °C in Dulbecco's modified Eagle medium (Nissui) containing 10% fetal bovine serum, 4.5 mg ml⁻¹ D-glucose, 200 μ g ml⁻¹ streptomycin, 100 units ml⁻¹ penicillin and 250 μ g ml⁻¹ G418 sulphate, and collected in phosphate-buffered saline containing 2.7 mM EDTA. Manipulations were then performed at 4 °C. Cells were washed with buffer P (50 mM HEPES (pH 6.6 at 4 °C), 140 mM NaCl, 3 mM MgCl₂, 1 mM DTT, 50 kallikrein-inhibitor units ml⁻¹ aprotinin, 4 μ g ml⁻¹ leupeptin) containing 2.7 mM EDTA, then homogenized with buffer P containing 250 mM sucrose. The suspension was layered on buffer P supplemented with 1.17 M sucrose, and centrifuged at 113,000g for 1 h. The membrane fraction sedimenting between the two sucrose layers was collected²¹ and washed twice with buffer P. The procedures described below were carried out under dim red light (>660 nm). The membrane fraction



containing 3 mg protein in 0.8 ml buffer P was mixed with 2 μ l ethanol solution of 1 mM 11-*cis*-retinal, incubated for 30 min, and centrifuged. Proteins in the regenerated membranes thus precipitated were solubilized with 0.2 ml of buffer P containing 0.75% CHAPS, 1 mg ml⁻¹ L- α -phosphatidylcholine from fresh egg yolk (type XI-E; Sigma), and 10 μ M 11-*cis*-retinal. This mixture was spun and the clear supernatant was used for spectrophotometry.

red-sensitive pigment is cooperatively involved in the regulation, and that a combination of blue-sensitive pinopsin and the red-sensitive pigment ($\lambda_{\max}$ 571 nm)¹⁷ would yield the broad action spectrum. In addition to the results from northern blot analysis (Fig. 2b), immunohistochemistry on quail pineal gland also supports the existence of at least two types of photoreceptors^{7,9}. Alternatively, an unidentified green-sensitive pigment (undetectable with chicken green or rhodopsin probes under the conditions used for Fig. 2b) may acutely inhibit *N*-acetyltransferase activity, with pinopsin contributing to the phase-shifting effect of light on the pacemaker. Identification of the steps in the pinopsin cascade, including a photo-signal-transducing G protein and its downstream effector, should indicate whether pinopsin triggers either or both of the two pathways, and help to clarify the molecular mechanism of the circadian pacemaker. □

Received 1 August; accepted 22 September 1994.

1. Binkley, S. A., Riebmman, J. B. & Reilly, K. B. *Science* **202**, 1198–1201 (1978).
2. Kasal, C. A., Menaker, M. & Prez-Polo, J. R. *Science* **203**, 656–658 (1979).
3. Deguchi, T. *Science* **203**, 1245–1247 (1979).
4. Deguchi, T. *Nature* **290**, 706–707 (1981).

5. Wallingford, J. C. & Zats, M. *Exp. Eye Res.* **46**, 909–918 (1988).
6. Takahashi, J. S., Murakami, N., Nikaïdo, S. S., Pratt, B. L. & Robertson, L. M. *Rec. Progr. Horm. Res.* **45**, 279–352 (1989).
7. Foster, R. G., Schalken, J. J., Timmers, A. M. & DeGrip, W. J. *J. comp. Physiol. A* **165**, 553–563 (1989).
8. Araki, M., Fukada, Y., Shichida, Y., Yoshizawa, T. & Tokunaga, F. *Dev. Brain Res.* **65**, 85–92 (1992).
9. Masuda, H. et al. *Tiss. Cell* **26**, 101–113 (1994).
10. van Veen, T. et al. *Proc. natn. Acad. Sci. U.S.A.* **83**, 912–916 (1986).
11. Dryer, S. E. & Henderson, D. *Nature* **353**, 756–758 (1991).
12. Kuwata, O. et al. *FEBS Lett.* **272**, 128–132 (1990).
13. Okano, T., Kojima, D., Fukada, Y., Shichida, Y. & Yoshizawa, T. *Proc. natn. Acad. Sci. U.S.A.* **89**, 5932–5936 (1992).
14. Saitou, N. & Nei, M. *Molec. Biol. Evol.* **4**, 406–425 (1987).
15. Kojima, D. et al. *Proc. natn. Acad. Sci. U.S.A.* **89**, 6941–6945 (1992).
16. Sun, J.-H., Reiter, R. J., Mata, N. L. & Tsien, A. T. *C. Neurosci. Lett.* **133**, 97–99 (1991).
17. Okano, T., Fukada, Y., Artamonov, I. D. & Yoshizawa, T. *Biochemistry* **28**, 8848–8856 (1989).
18. Zats, M. & Mullen, D. A. *Brain Res.* **453**, 63–71 (1988).
19. Takao, M., Yasui, A. & Tokunaga, F. *Vision Res.* **28**, 471–480 (1988).
20. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* 2nd edn (Cold Spring Harbor Laboratory Press, New York, 1989).
21. Nathans, J. *Biochemistry* **29**, 9746–9752 (1990).

ACKNOWLEDGEMENTS. The GenBank accession number for pinopsin cDNA reported here is U15762. We thank D. Kojima, K. Yamazaki and Y. Takanaka for their assistance in the experiments and T. Miyata and N. Iwabe for discussion. This work was supported in part by Grants-in-Aid from the Japanese Ministry of Education, Science and Culture, and by research grants of Toray Science Foundation and SUNBOR. T.O. was supported by Fellowships of the Japan Society for the Promotion of Science for Young Scientists.

Molecular determinants of voltage-dependent inactivation in calcium channels

Ji-Fang Zhang, Patrick T. Ellinor, Richard W. Aldrich* & Richard W. Tsien†

Department of Molecular and Cellular Physiology,
* Howard Hughes Medical Institute, Stanford University Medical
Center, Stanford, California 94305, USA

VOLTAGE-DEPENDENT Ca^{2+} channels respond to membrane depolarization by conformational changes that control channel opening and eventual closing by inactivation^{1–3}. The kinetics of inactivation differ considerably between types of Ca^{2+} channels^{1–8} and are important in determining the amount of Ca^{2+} entry during electrical activity and its resulting impact on diverse cellular events³. The most intensively characterized forms of inactivation in potassium^{9,10} and sodium channels^{11–13} involve pore block by a tethered plug¹⁴. In contrast, little is known about the molecular basis of Ca^{2+} -channel inactivation. We studied the molecular mechanism of inactivation of voltage-gated calcium channels by making chimaeras from channels with different inactivation rates. We report here that the amino acids responsible for the kinetic differences are localized to membrane-spanning segment S6 of the first repeat of the α_1 subunit (IS6), and to putative extracellular and cytoplasmic domains flanking IS6. Involvement of this region in Ca^{2+} -channel inactivation was unexpected and raises interesting comparisons with Na^+ channels, where the III–IV loop is a critical structural determinant. Ca^{2+} -channel inactivation has some features that resemble C-type inactivation of potassium channels.

We studied calcium channel inactivation using three different Ca^{2+} -channel α_1 subunits^{8,15,16}: doe-1, BI (α_{1A}) and cardiac L-type (Card). These α_1 subunits displayed distinctive inactivation properties when coexpressed with α_2 and β_1 subunits in *Xenopus* oocytes (Fig. 1). Doe-1 inactivated considerably faster than the other channels whereas BI either inactivated faster than Card (–10 mV, Fig. 1a) or at a similar rate (+20 mV, Fig. 1b). With Ba^{2+} as the charge carrier to minimize any Ca^{2+} -dependent inactivation^{3,17}, the rate of inactivation increased monotonically with depolarization for all three channels, saturating at strongly

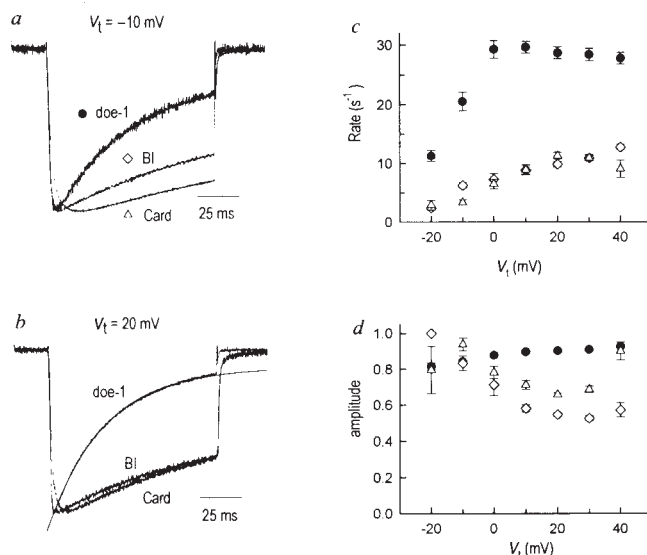


FIG. 1 Comparison of kinetic properties of three calcium-channel α_1 subunits. a and b, Examples of current traces for doe-1, BI and cardiac L-type Ca^{2+} channels, elicited by step depolarizations from a holding potential of –90 mV to test potentials (V_t) as indicated. All current traces are scaled to the same peak amplitude to allow comparison of kinetics. c, Rate of exponential decay (s^{-1}) plotted against voltage. d, Amplitude of the inactivating component plotted against voltage. Data points represent mean $\pm$ s.e.m. ($n = 5-8$).

METHODS. Complementary RNA for α_1 subunits was synthesized *in vitro* using T7 or SP6 polymerase and the pDoe-1, pSPCBI-2, pSPCA1 and pCaB1 plasmids. All mutant channels (chimaeras and point mutations) were coexpressed with α_2 and β_1 subunits. Subunit cRNAs were injected in roughly equimolar ratios using $0.1 \mu\text{g ml}^{-1}$ α_1 , $0.1 \mu\text{g ml}^{-1}$ α_2 and $0.03 \mu\text{g ml}^{-1}$ β_1 ; ~ 50 nl was injected per cell. Currents were recorded with a standard two-microelectrode voltage clamp (Warner OC-725A). The bath solution was chloride free, containing (in mM): $\text{Ba}(\text{OH})_2$, 5; TEA-OH, 85; KOH, 2; HEPES, 5; pH adjusted to 7.4 with methane sulphonic acid. Data were sampled at 10 kHz and filtered at 2 kHz. Leak and capacitive currents were subtracted off-line by a P/4 protocol. Voltage pulses were delivered every 15 s. Rate constants represent the reciprocal of the exponential time constant (τ_h), obtained by fitting the normalized current trace to the function $y = A \exp(-t/\tau_h) + B$ (smooth curve in b). The ratio $A/(A+B)$ is the amplitude of the decay, as plotted in d. All experiments were done at room temperature ($\sim 23^\circ\text{C}$).

† To whom correspondence should be addressed.

positive potentials (Fig. 1c). The maximal rate for *doe-1* was two- to threefold higher than for either BI or Card. For *doe-1*, the fractional amplitude of the decaying component remained near 0.8 at all membrane potentials, whereas it fell considerably at strong depolarizations for BI or Card (Fig. 1d).

The intracellular loop linking motifs III and IV may be an important determinant of inactivation in sodium channels^{11,13}, which resemble Ca^{2+} channels in architecture¹³. Thus, variations in the III-IV loop of Ca^{2+} channels seemed a likely source of the differences in inactivation behaviour. To test this, we replaced the III-IV loop of a host α_{1A} subunit with that from another α_1 subunit (Fig. 2a). The chimera containing the III-IV loop of the rapidly inactivating channel *doe-1* (BD3-4) inactivated much more slowly than *doe-1* and only slightly faster than BI (Fig. 2a, b). The slight speeding in inactivation of BD3-4 relative to BI was no greater than that produced by introducing the III-IV loop of Card into BI (BC3-4, Fig. 2c), or vice versa (data not shown). The absence of any dramatic change in inactivation with III-IV loop replacement was not anticipated from previous work on Na^+ channels.

To determine which parts of the channel molecule are responsible for the differences in inactivation kinetics, we constructed half-half chimaeras. In the chimera that combined the N-terminal half of *doe-1* with the C-terminal half of BI (DBHH: Fig. 2a), inactivation was much faster than in wild-type BI, with a rate approaching that of *doe-1* itself (Fig. 2a, d) and with corresponding changes in the voltage dependence of the fractional decay. DBHH3-4, which consisted of DBHH with a III-

IV loop from *doe-1*, behaved like DBHH itself (Fig. 2a). In contrast to DBHH, a chimera whose N-terminal half was Card (CBHH) showed less pronounced inactivation than either Card or BI (Fig. 2a, e), suggesting that the rapid inactivation of DBHH was not simply a result of a structural mismatch.

To narrow the region critical for inactivation to a smaller domain within the N-terminal half, additional chimaeras were made between *doe-1* and BI. These are presented as complementary pairs in which the contribution of *doe-1* lies on one side or the other of a common boundary (Fig. 3a). In DB6 and DB9, for example, the boundary lies near the end of the I-II loop. In DB6 (*doe-1* segment N-terminal to boundary), inactivation was much faster and the amplitude of the inactivating component was considerably greater than for BI, whereas the complementary chimera DB9 inactivated at a rate quite close to that for BI (Fig. 3a). Thus, motif II does not appear to play a critical role in determining the differences in the inactivation process.

In DB4 and DB10, the chimaeric boundary was positioned C-terminal to the pore-forming region in IH5. DB4, which consisted of *doe-1* up to this boundary, inactivated with rate constant and fractional decay very similar to those of BI, whereas the complementary chimera DB10 inactivated much more rapidly and completely than BI ($21.9 \pm 2.3 \text{ s}^{-1}$ versus $9.8 \pm 0.5 \text{ s}^{-1}$, $73 \pm 1.0\%$ versus $55 \pm 0.8\%$, respectively). Thus, inactivation depends critically on a segment <200 residues long, extending from the middle of IH5 to the beginning of motif II, which represents the difference between DB9 and DB10 or between DB4 and DB6.

FIG. 2 Assessment of the importance of the III-IV loop and the N-terminal half of the α_1 subunit in determining the rate of inactivation. a, Schematic representation of III-IV loop chimaeras and half-and-half chimaeras, and comparison of inactivation rate constants at +20 mV ($n=5-8$). In this and other figures, boxes represent 25th and 75th percentiles, the line within the box gives the median, the error bars indicate the 10th and 90th percentiles, and the dots are 5th and 95th percentiles. b and d, Representative current traces for BI/*doe-1* chimaeras along with wild-type channels ($V_t = +20 \text{ mV}$). All current traces were scaled to the same peak amplitude to emphasize possible differences in kinetics. c and e, Representative current traces for BI/Card chimaeras.

METHODS. All chimaeras and point mutants were constructed using standard molecular biological techniques. Two unique and silent restriction sites for *Bst*1107I (4,796) and *Bst*BI (5,022) were introduced into BI. To make chimera BDI, a fragment of *doe-1* generated by polymerase chain reaction (PCR) was ligated into this construct at the *Bst*1107I and *Bst*BI sites. Chimera BC3-4 was constructed by generating a fragment of CARD by the PCR, and introducing it into BI at the *Xcm*I (4,783) and *Msc*I (5,038) sites. Half-half chimaeras (DBHH and CBHH) were made as illustrated. Construct CBHH (also known as BC2) was generously provided by Y. Mori and T. Tanabe. Chimera DBHH was constructed as follows. (1) A PCR fragment of *doe-1* from 2,744 to 2,802 was ligated into BI at the *Sal*I (polylinker) and *Rsr*II (3,212) sites; (2) the 2.7-kb *Sal*I fragment of *doe-1* (polylinker-2,744) was ligated into the construct from (1) at the *Sal*I site. Chimera DBHH3-4 was constructed from BD3-4 in the same way that DBHH was made from BI. Construct CBHH3-4 was made by ligating the *Nhe*I (3,542) and *Xba*I (polylinker) fragment of BC3-4 into the *Nhe*I and *Xba*I sites of CBHH. See Fig. 1 for recording conditions and data analysis. The effects of chimaeras were confirmed with at least two independent clones.

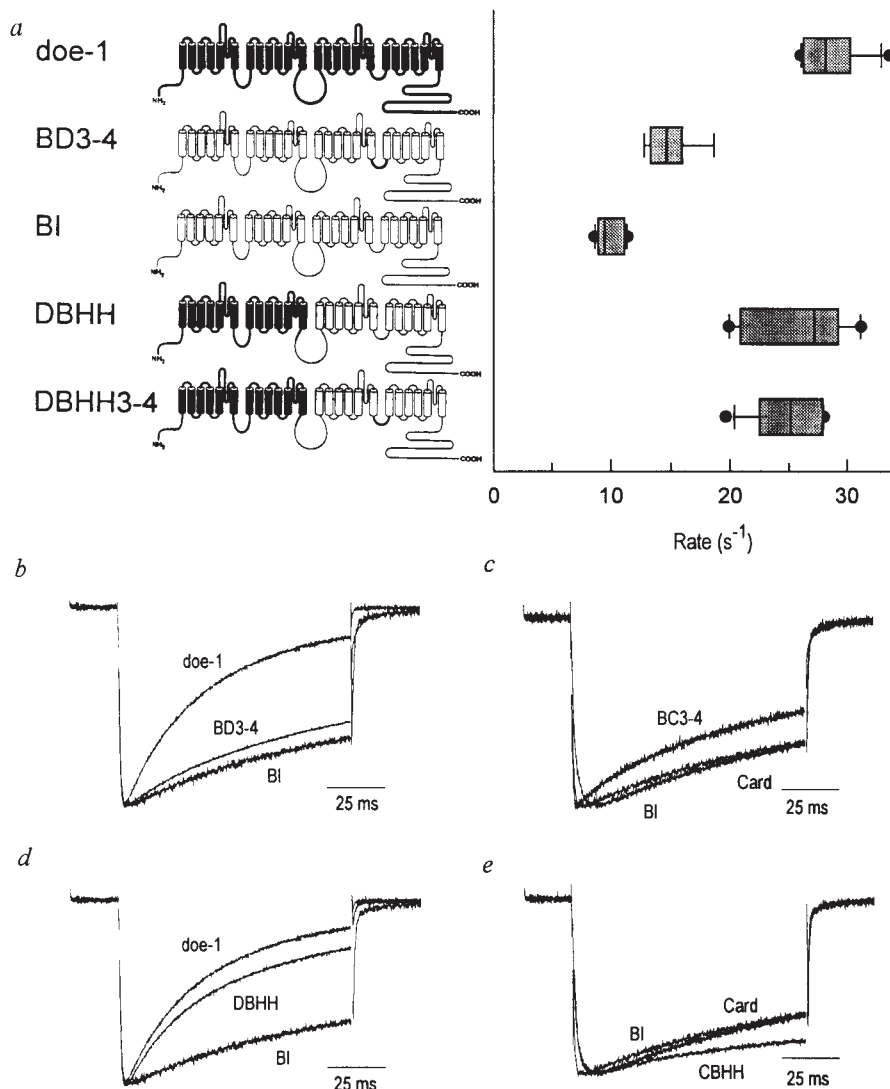
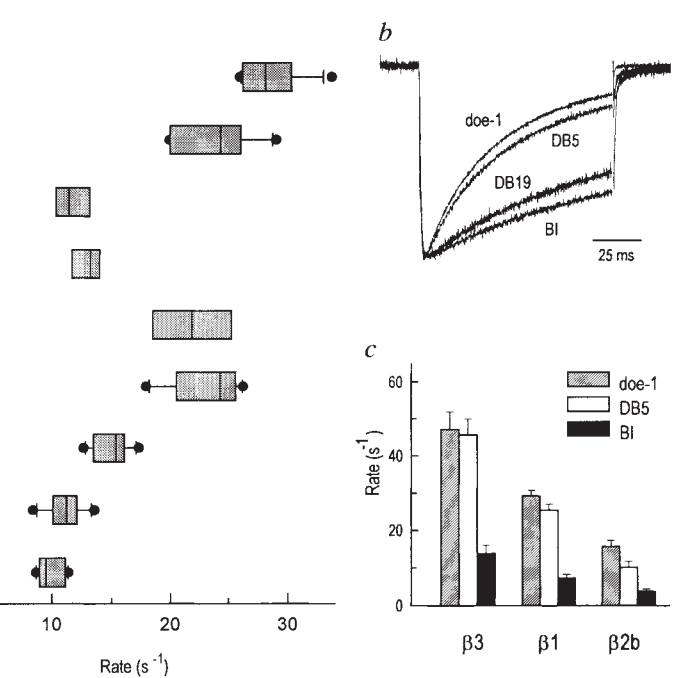


FIG. 3 Localization of structural determinants involved in inactivation. **a**, Chimaeric constructs and their inactivation rate constants: *doe-1* represented by filled cylinders and bold lines, BI by unfilled cylinders and thin lines. For the chimaeras, only the first half of the construct is shown. Box plots show rate constants of decay of currents generated by the chimaeras and wild-type channels ($V_t = +20$ mV, $n = 4-8$). **b**, Scaled current traces for *doe-1*, BI, DB5 and DB19 ($V_t = +20$ mV). **c**, Comparison of inactivation rate constants for *doe-1*, BI and DB5 co-expressed with α_2 and three different β -subunits ($V_t = 0$ mV, $n = 4-9$).

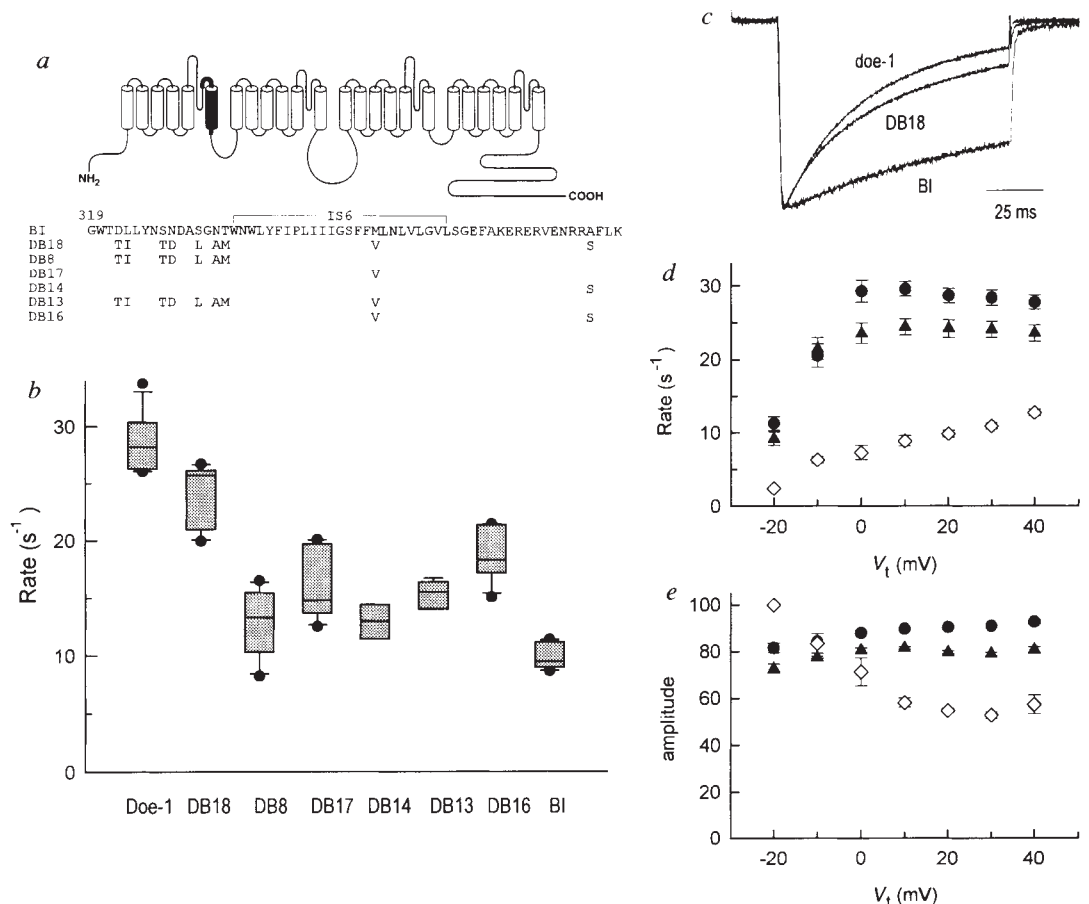
METHODS. Construct BI-5 is a modified version of BI which contains a silent *NsiI* site introduced at 1,232 and a deletion of the *BamHI* fragment (7,612-3'end). Chimaeras between BI and *doe-1* in the first and second motifs were constructed by introducing a *doe-1* fragment into BI-5 at the following sites: DB6, *Sall* (polylinker) and *XhoI* (1,688); DB9, *XhoI* (1,688) and *HindIII* (2,502); DB4, *Sall* (polylinker) and *NsiI* (1,232); DB10, *NsiI* (1,232) and *HindIII* (2,502); DB5, *NsiI* (1,232) and *XhoI* (1,688); DB7, *SacI* (1,455) and *XhoI* (1,688). Chimaera DB5/3-4 was assembled by ligating the *NheI* (3,542) to *BamHI* (polylinker) fragment of BDI into the *NheI*



and *BamHI* sites of DB5. For DB19, a BI fragment was ligated into the *doe-1* half of DB1 at *NsiI* (948) and *BstXI* (1,388). For most constructs, at least two independent clones were used for current measurements.

FIG. 4 Localization of amino-acid residues involved in inactivation. **a**, Amino-acid sequence of wild-type BI (first line, residues 319 to 378), and modifications of this sequence according to that of *doe-1*. DB18 (second line) includes all amino-acid differences between BI and *doe-1* within this stretch of the α_1 subunit. In this region of other α_1 subunits, there is a general correlation between their similarity to *doe-1* and their rapidity of inactivation^{30,31}. None of the BI $\rightarrow$ *doe-1* replacements altered consensus sites for protein phosphorylation. **b**, Box plots of the rate constants for mutations illustrated in **a** ($V_t = +20$ mV). **c**, Representative normalized traces for *doe-1*, DB18 and BI. **d** and **e**, Voltage-dependence of the exponential rate constant and amplitude of decay for *doe-1* (●), DB18 (△) and BI (◇). Data are mean $\pm$ s.e.m. ($n = 4-8$). The behaviour of DB18 was confirmed with two independent clones.

METHODS. Individual amino-acid changes were made using the mega-primer PCR method. Complementary DNA fragments containing the point mutations were introduced into motif I at the *NsiI* (1,232) and *XhoI* (1,688) sites of BI-5 as illustrated in **a**. All point mutations were confirmed by sequencing.



As a further test, this specific segment of *doe-1* was inserted into a BI background to yield chimaera DB5 (Fig. 3a). As expected, DB5 closely resembled *doe-1* in its current waveform (Fig. 3b), and in the voltage-dependence of the rate and degree of inactivation. In contrast, DB19, a chimaera complementary to DB5 in that the contributions of *doe-1* and BI within the first two motifs were switched (Fig. 3a), inactivated at close to the same rate as BI (Fig. 3a, b). DB5 and DB19 had significant differences in the rate of inactivation ($23.15 \pm 1.32 \text{ s}^{-1}$ versus $14.93 \pm 0.65 \text{ s}^{-1}$) as well as its extent ($79 \pm 2.8\%$ versus $66 \pm 1.2\%$, respectively; $n = 6$ or 7). These results substantiate the importance of residues within the <200 amino-acid segment near IS6.

Although β -subunits significantly modify Ca^{2+} -channel kinetics^{18,19}, the effect of substituting a segment of *doe-1* for BI was a general effect, not restricted to a particular β -subunit (Fig. 3c). Whether DB5, *doe-1* or BI were expressed with β_1 , β_{2b} or β_3 (the α_2 subunit remaining unchanged), the chimaera always inactivated faster than BI. For DB5, β -subunits ranked $\beta_3 > \beta_1 > \beta_{2b}$ with respect to speed of inactivation, just as for BI, *doe-1* or Card channels⁸.

The chimaera DB5 encompasses an 18 amino-acid segment within the I-II loop that contributes to interactions of the α_1 subunit with the β -subunit²⁰. When this segment and adjoining residues within the I-II loop of BI were replaced with corresponding residues of *doe-1* (chimaera DB7, Fig. 3a), inactivation was similar to that of BI. Evidently, the region responsible at least in part for interactions with the β -subunit²⁰ is not critical in determining the differences in inactivation kinetics between *doe-1* and BI. As further confirmation, behaviour such as that of DB5 was seen with a more restricted chimaeric construct (DB18) in which the 18 amino-acid segment was left as in BI. The close resemblance of DB18 to *doe-1* rather than BI is seen in current traces (Fig. 4c) and in the voltage-dependence of the rate or extent of inactivation (Fig. 4d, e).

Chimaera DB18 encompasses only nine positions of amino-acid difference between *doe-1* and BI (Fig. 4a). These positions were further subdivided into (1) a cluster of seven positions located at the end of the extracellular loop linking IH5 and IS6; (2) one position in the putative membrane-spanning region IS6; and (3) one position appearing just C-terminal to IS6, in a putative cytoplasmic region (Fig. 4a). Each of these groupings was mutated individually, yielding the constructs DB8, DB17 and DB14 (Fig. 4a). None of these constructs rivalled the rate of inactivation of *doe-1*, although DB17 (350M $\rightarrow$ V) came the closest (Fig. 4b). DB13 and DB16 are additional constructs which combine two groups of positions (Fig. 4a). Interestingly, DB16, the combination of two single-point mutations (350M $\rightarrow$ V, 374A $\rightarrow$ S), has the more rapid inactivation (Fig. 4b).

We have identified structural determinants of Ca^{2+} -channel inactivation which are different from those previously shown to control N-type inactivation of K^+ channels and fast inactivation in Na^+ channels^{9,13}. The rate or extent of Ca^{2+} channel inactivation was hardly affected by the source of the N-terminal cytoplasmic region, unlike N-type inactivation, or by the origin of the III-IV loop, contrary to what might be expected by analogy to fast Na^+ -channel inactivation. In contrast, transplantation of nine or fewer amino acids near IS6 was sufficient to switch the inactivation behaviour from BI-like to *doe-1*-like (Fig. 4). Some additional involvement of the N-terminal or the III-IV loop cannot be ruled out, although this would have to involve residues that are identical between BI and *doe-1*. Nor can we exclude the possibility that IS6 contributes to a receptor site for a tethered plug. Nonetheless, some features of Ca^{2+} -channel inactivation are reminiscent of another form of inactivation in K^+ channels, called C-type²¹⁻²³, which depends strongly on residues in domains H5 and S6, situated close to the ion-conducting pore^{24,25}. It seems possible that Na^+ channels are also capable of a form of C-type inactivation, which may be unmasked by eliminating rapid inactivation involving the III-IV loop^{11,13}. This would help explain why Na^+ -channel inactivation may be

affected by substitutions in extracellular loops and membrane-spanning segments, as in periodic paralysis and paramyotonia^{26,27}, or by pharmacological reagents that interact with extracellular regions^{3,28}, including α -scorpion toxin, whose effects on Na^+ -channel inactivation are prevented by antibodies to the IH5-IS6 linker²⁹. We suggest that the diversity of inactivation properties among Ca^{2+} , K^+ and Na^+ channels might be explained in terms of the same few basic mechanisms, working in different combinations. □

Received 20 May; accepted 27 September 1994.

1. Carbone, E. & Swandulla, D. *Prog. Biophys. molec. Biol.* **54**, 31–58 (1989).
2. Reuter, H. A. *Rev. Physiol.* **46**, 473–484 (1984).
3. Hille, B. *Ionic Channels of Excitable Membranes* (Sinauer, Sunderland, MA, 1992).
4. Armstrong, C. M. & Matteson, D. R. *Science* **227**, 65–67 (1985).
5. Nowicky, M. C., Fox, A. P. & Tsien, R. W. *Nature* **316**, 440–443 (1985).
6. Bean, B. A. *Rev. Physiol.* **51**, 367–384 (1989).
7. Llinas, R. et al. *Trends Neurosci.* **15**, 351–355 (1992).
8. Ellinor, P. T. et al. *Nature* **363**, 455–458 (1993).
9. Hoshi, T., Zagotta, W. N. & Aldrich, R. W. *Science* **250**, 533–538 (1990).
10. Demo, S. D. & Yellen, G. *Neuron* **7**, 743–753 (1991).
11. Stuhmer, W. et al. *Nature* **339**, 597–603 (1989).
12. Vassilev, P., Scheuer, T. & Catterall, W. A. *Proc. natn. Acad. Sci. U.S.A.* **86**, 8147–8151 (1989).
13. Catterall, W. A. *Trends Neurosci.* **16**, 500–506 (1993).
14. Armstrong, C. M. & Bezanilla, F. *J. gen. Physiol.* **70**, 567–590 (1977).
15. Mori, Y. et al. *Nature* **350**, 398–402 (1991).
16. Mikami, A. et al. *Nature* **340**, 230–233 (1989).
17. Eckert, R. & Chad, J. D. *Prog. Biophys. molec. Biol.* **44**, 215–267 (1984).
18. Lacerda, A. E. et al. *Nature* **352**, 527–530 (1991).
19. Singer, D. et al. *Science* **253**, 1553–1557 (1991).
20. Pragnell, M. et al. *Nature* **368**, 67–70 (1994).
21. Hoshi, T., Zagotta, W. N. & Aldrich, R. W. *Neuron* **7**, 547–556 (1991).
22. Kirsch, G. E. et al. *Biophys. J.* **62**, 136–143 (1992).
23. Choi, K. L., Aldrich, R. W. & Yellen, G. *Proc. natn. Acad. Sci. U.S.A.* **88**, 5092–5095 (1991).
24. Lopez-Barneo, J. et al. *Receptors Channels* **1**, 61–71 (1993).
25. Yellen, G. et al. *Biophys. J.* **66**, 1068–1075 (1994).
26. Rojas, C. V. et al. *Nature* **354**, 387–389 (1991).
27. Cannon, S. C. & Strittmatter, S. M. *Neuron* **10**, 317–326 (1993).
28. Strichartz, G., Rando, T. & Wang, G. K. A. *Rev. Neurosci.* **10**, 237–267 (1987).
29. Thomsen, W. J. & Catterall, W. A. *Proc. natn. Acad. Sci. U.S.A.* **86**, 10161–10165 (1989).
30. Williams, M. E. et al. *Science* **257**, 390–395 (1992).
31. Soong, T. W. et al. *Science* **260**, 1133–1136 (1993).

ACKNOWLEDGEMENTS. We thank T. Tanabe and Y. Mori for cardiac and BI α_1 and skeletal α_2 subunits as well as chimaera BC2, F. Hofmann for β_{2b} and β_3 subunits and S. Siegelbaum, W. Zagotta, D. Gross, P. Drain, A. Randall, W. Sather and G. Liu for comments on the manuscript. This work was supported by grants from the NIH (R.W.A. and R.W.T.), a postdoctoral fellowship from the American Heart Association (California Affiliate) (J.F.Z.) and an MSTP fellowship (P.T.E.). R.W.A. is an investigator with the Howard Hughes Medical Institute.

T-cell apoptosis detected *in situ* during positive and negative selection in the thymus

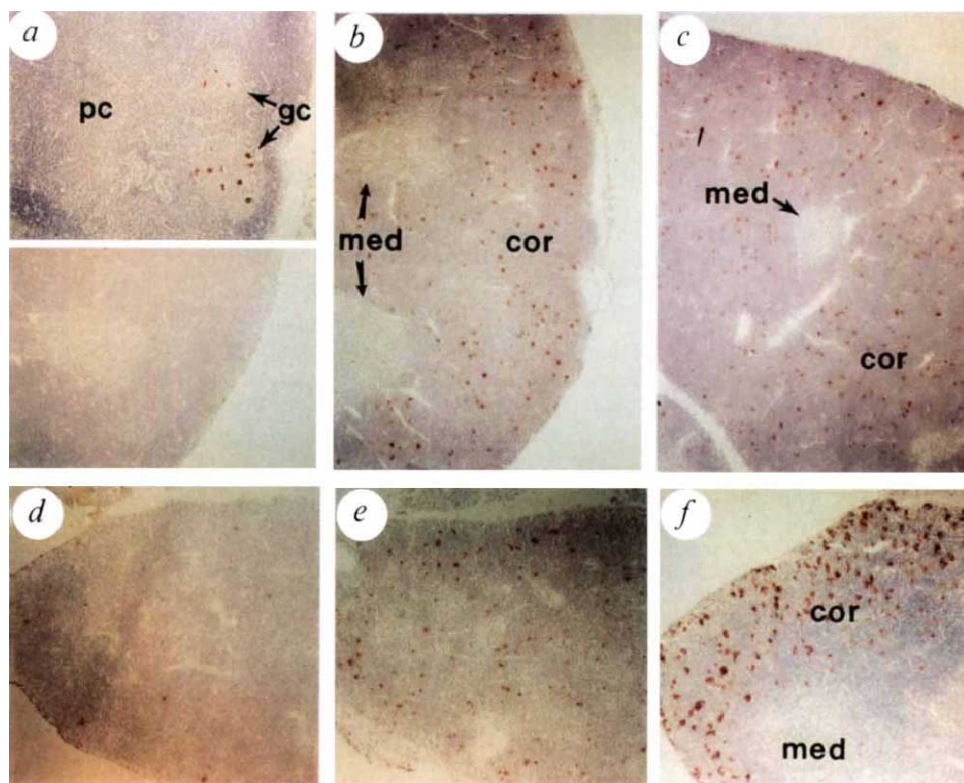
Charles D. Surh & Jonathan Sprent

Department of Immunology IMM4, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, California 92037, USA

BECAUSE of positive and negative selection to molecules of the major histocompatibility complex (MHC)¹, only a small proportion of the massive numbers of T cells generated in the thymus are selected for export^{2,3}. Immature thymocytes have a rapid turnover², and it has long been assumed that most thymocytes die *in situ*^{4,5}, presumably from apoptosis⁶. This has yet to be proved, however, and conventional staining techniques have shown only minimal evidence of cell death in the normal thymus^{7,8}. Using a method for detecting cells with DNA strand breaks, we now present direct evidence for apoptosis in the normal thymus. In sections of thymus from adult mice, apoptotic cells are scattered throughout the cortex and are engulfed locally by F4/80⁺ macrophages. Apoptosis in the thymic cortex is not reduced in MHC-deficient mice, which suggests that T-cell death is primarily a reflection of lack of positive selection rather than negative selection. Direct evidence for apoptosis due to negative selection was obtained by crossing a V β 5 transgenic line⁹ to I-E⁺ and I-E⁻ mice: I-E⁺ mice

FIG. 1 *In situ* detection of apoptotic cells in cryostat sections of thymus with a modified TUNEL method ($\times 45$). *a*, Top, normal C57BL/6 (B6) lymph nodes: staining is prominent in germinal centres (gc) but is sparse in the paracortex (pc) and the mantle zones surrounding germinal centres. Bottom, B6 thymus stained in the absence of TdT (negative control): staining is undetectable. *b*, Normal B6 thymus: stained cells are scattered throughout the cortex (cor) but are rare in the medulla (med). *c*, Thymus from MHC class I + II-deficient mouse: stained cells are present in the cortex at about the same frequency as in the normal thymus; the medulla is small, reflecting the paucity of mature T cells in MHC⁻ mice. *d*, Gestation day-18 fetal B6 thymus: stained cells are rare. *e*, Day-2 postnatal B6 thymus: stained cells are as common as in the adult thymus. *f*, Thymus from B6 mouse injected with anti-CD3 monoclonal antibody 16 h before: stained cells are numerous in the cortex and are much larger than in the normal thymus (compare with *b*).

METHODS. Fresh lymphoid tissues were immersed in OCT compound (Miles), rapidly frozen in liquid nitrogen, and 5–6- μ m sections were cut in a cryostat. Sections were air-dried, fixed in acetone and incubated with 0.01% H₂O₂ diluted in Tris-buffered saline (pH 7.4) to neutralize endogenous peroxidases. To detect cells with DNA fragmentation, we modified the previously described TUNEL method¹¹ as follows. Tissue sections were washed once with TdT buffer (0.5 M cacodylate, pH 6.8, 1 mM CoCl₂, 0.5 mM DTT, 0.05% BSA, 0.15 M NaCl) and then incubated with 2–4 μ M digoxigenin-conjugated dUTP (Boehringer Mannheim) and 5–10 U TdT (Promega) in 25 μ l TdT buffer. After washing in Tris-saline, sections were incubated with 5 μ g ml⁻¹ sheep Fab antibody fragments specific for digoxigenin (Boehringer Mannheim) in Tris-saline, washed and then incubated with 10 μ g ml⁻¹ horseradish-peroxidase (HRP)-conjugated antibody specific for sheep immunoglobulin (Jackson ImmunoResearch). Bound HRP was detected with the substrate 3-amino-9-ethylcarbazole (0.1 mg ml⁻¹ in 0.17 M sodium acetate, pH 5.2, plus



0.01% H₂O₂). Sections were lightly counterstained with Meyer's haematoxylin and photographed under a Zeiss Axioscope microscope. Mice deficient in expression of both MHC class I and II molecules were obtained from F₂ breeding between B6-background β_2 -microglobulin-deficient²⁴ and I-A-deficient mice²⁵. Normal adult mice were analysed at 6 weeks old and MHC-deficient mice were analysed at 4 weeks old. For anti-CD3 antibody-induced cell death, mice were injected i.p. with 50 μ g purified 2C11 antibody in PBS (Pharmingen) and analysed 16 h later.

are known to eliminate V β 5⁺ T cells in the thymus whereas I-E⁻ mice do not¹⁰. In marked contrast to I-E⁻ mice, the medulla of I-E⁺ V β 5 transgenic mice contains dense aggregates of apoptotic cells; these cells are engulfed by a distinct population of F4/80⁺ MAC-3⁺ macrophages. Negative selection of V β 5⁺ cells is thus restricted to the medulla.

Terminal deoxynucleotidyl transferase (TdT)-mediated dUTP-biotin nick end-labelling (TUNEL)¹¹ detects DNA strand breaks in cells undergoing apoptosis. Thus, unlike normal cells, nuclei of apoptotic cells incorporate exogenous nucleotides (dUTP-biotin) in the presence of TdT. When amplified with a three-step staining procedure, the TUNEL method readily detects apoptotic cells in cryostat sections of normal lymphoid tissues (Fig. 1). In mouse spleen and lymph nodes, staining is limited to isolated cells in germinal centres, a known site for apoptotic cells¹² (Fig. 1*a*, top). In the thymus, however, stained cells are quite prominent. In normal adult thymus, stained cells are sparse in the thymic medulla but are found scattered throughout the cortex (0.5–1.0% of mononuclear cells) (Fig. 1*b*); no staining occurs in the absence of exogenous TdT (Fig. 1*a*, bottom). To exclude artefact, we examined the thymus of fetal mice at gestation day 18, namely at a stage when positive selection is just beginning¹ and cell death from lack of positive selection is presumably very limited. Significantly, TUNEL staining of the fetal thymus is quite low (Fig. 1*d*). By day 2 after birth, however, staining is as prominent as in the adult thymus (Fig. 1*e*). Based on these findings, we conclude that staining with the TUNEL technique is not artefactual and is specific for apoptotic cells.

As very few cortical thymocytes undergo positive selection¹, it can be argued that cell death in the cortex is primarily a

reflection of 'neglect' (lack of positive selection) rather than negative selection. If so, one would expect the incidence of cell death in the cortex to be undiminished in MHC-deficient mice¹³; here, the lack of MHC molecules precludes both positive and negative selection, and cell death in the thymus is presumably due solely to neglect. As shown in Fig. 1*c*, the frequency of apoptotic cells in the cortex of mice deficient in both MHC I and II is about the same as in normal mice.

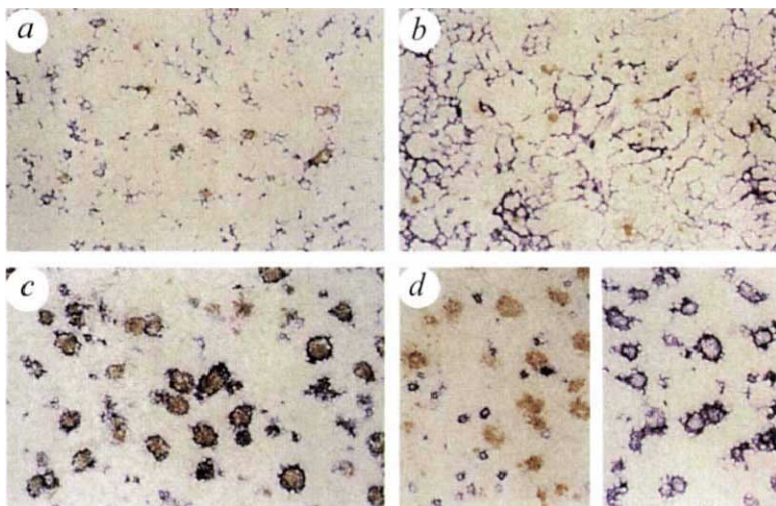
The fate of apoptotic thymocytes is shown in Fig. 2. Double staining for thymic epithelial cells and macrophages (stained blue) reveals that most of the apoptotic cells (red) in the cortex of the normal thymus lie apart from the network of thymic epithelial cells (Fig. 2*b*) and are situated within small F4/80⁺ macrophages (Fig. 2*a*). This finding agrees with reports that apoptotic cells are rapidly engulfed by macrophages^{14,15} and that in the electron microscope, many of the macrophages in the thymus contain apoptotic bodies¹⁶.

Cortical thymocytes are highly sensitive to corticosteroids, irradiation and anti-CD3 monoclonal antibody^{17–20}. As demonstrated by anti-CD3 antibody (Fig. 1*f*), these agents cause a rapid, marked increase in apoptotic cells in the cortex; the apoptotic cells aggregate in large clusters, and double staining reveals that these foci of apoptotic cells are situated within massively distended F4/80⁺ macrophages (Fig. 2*c*; compare with Fig. 2*a*). The apoptotic foci in macrophages are not artefactual because staining is TdT-dependent (Fig. 2*d*, right) and, at least for corticosteroid treatment, the aggregates of apoptotic cells do not include CD25⁺ cells (a subset of CD4⁺ 8⁻ cells expressing IL-2R; ref. 1) (Fig. 2*d*, left).

Of the three treatments tested, irradiation causes the earliest onset of apoptosis. Apoptotic cells in the cortex are increased

FIG. 2 Ingestion of apoptotic cells in the thymus by macrophages ($\times 164$). *a*, Cortex of normal B6 thymus double-stained for apoptotic cells (red) and for F4/80, a macrophage marker (blue): nearly all apoptotic cells are situated inside macrophages. *b*, Cortex of normal B6 thymus double-stained for apoptotic cells (red) and for 6C3, a marker for cortical epithelial cells: few if any of the apoptotic cells are in epithelial cells. *c*, *d*, Cortex of thymus from B6 mouse treated 4 h before with hydrocortisone; in *c*, double staining for apoptotic cells (red) and for F4/80 (blue) reveals that apoptotic cells are situated within enlarged macrophages; in *d*, left, double staining for apoptotic cells (red) and for CD25 (IL-2R), a marker for a subset of $CD4^+8^+$ cells⁴, reveals that apoptotic cells do not include CD25⁺ cells; in *d*, right, the same staining procedure as in *c*, but without TdT shows no red staining of macrophages, indicating that the staining for apoptosis in macrophages is not artefactual. All photographed at $\times 164$.

METHODS. After staining for apoptosis (Fig. 1 legend), thymus sections were incubated sequentially with (1) F4/80 (ref. 26), 6C3 (ref. 26) or anti-CD25 (7D4; ref. 27) rat antibodies; (2) sheep anti-digoxigenin antibody plus biotinylated mouse anti-rat immunoglobulin antibody (Jackson ImmunoResearch); (3) HRP-conjugated anti-sheep immunoglobulin plus alkaline phosphatase-conjugated streptavidin (Jackson). Sections were



then developed with alkaline phosphatase substrate Fast blue followed by HRP substrate 3-amino-9-ethylcarbazole as described²⁶.

within 1 h (but not 30 min) of exposure to low-dose irradiation (100 cGy) and are plentiful by 2 h (Fig. 3*a, b*). The F4/80⁺ macrophages ingesting the apoptotic cells are only slightly enlarged at 1 h but are conspicuously distended by 2–4 h (Fig. 3*b, c*). By 24 h, despite a 3–4-fold decrease in the total cellularity of the thymus, F4/80⁺ macrophages are no longer distended and apoptotic inclusions are scarce (Fig. 3*d*). Although nearly all apoptotic cells detected at 1–2 h post-irradiation are found in macrophages, appreciable numbers of 'free' (not macrophage-associated) apoptotic cells are apparent at 4 h (Fig. 3*c*), perhaps reflecting macrophage overloading. By fluorescence-activated cell-sorting analysis, however, these free (F4/80⁻) apoptotic cells account for <5% of cells in thymocyte suspensions (data not shown). When higher doses of radiation are used (600 cGy), apoptotic cells are prominent in cell suspensions at 4 h (40–60%) but are scarce by 24 h (data not shown).

Our results indicate that removal of apoptotic cells by thymic macrophages is extremely efficient and leads to rapid destruction of dying cells. The efficiency of macrophages in clearing apoptotic cells would explain the paradox that the presumed high rate of cell death in the normal thymus is not obvious in tissue sections: thymocytes are dying continuously in the cortex but are cleared so rapidly that cell death is inconspicuous.

Negative selection in the thymus can occur in the cortex²¹ but is thought to take place predominantly in the medulla through

contact with cells derived from bone marrow¹. Much of the data on negative selection have come from studies on V β -specific deletion of T cells to endogenous mouse mammary tumour virus (MMTV) superantigens^{1,10,22,23}. V β -specific deletion to MMTV superantigens occurs during the transition of thymocytes to mature T cells, but whether the cells die in the cortex or medulla is not clear^{22,23}. We examined this issue with the aid of V β 5 transgenic mice⁹. Because of their specificity for superantigens 8 and 9 of MMTV^{9,10}, mature V β 5⁺ T cells are generated in I-A⁺/I-E⁻ mice but are deleted in I-A⁺/I-E⁺ mice during late thymocyte differentiation. The TUNEL method reveals normal numbers of apoptotic cells in the cortex of both I-E⁺ and I-E⁻ V β 5 transgenic mice (Fig. 4*a, b*). As in normal mice, apoptotic cells are scarce in the medulla of I-E⁻ V β 5 transgenic mice (Fig. 4*b, d*). In striking contrast, the medulla and the cortico-medullary junction of I-E⁺ mice contain dense accumulations of apoptotic cells (Fig. 4*a, c*). Double staining for macrophage markers reveals that the apoptotic cells in the medulla and cortico-medullary junction are situated in MAC-3⁺ macrophages (Fig. 4*f*). But in contrast to the cortex, these macrophages are F4/80⁻ (Fig. 4*e*). Double staining for other markers suggests that the apoptotic cells are not engulfed by UEA-1⁺ medullary epithelial cells or B220⁺ B cells (not shown).

These data provide the first direct evidence for negative selection occurring in the medulla. Negative selection in this site

FIG. 3 Apoptotic cells in the cortical region of the thymus at different intervals after exposure to low-dose (100 cGy) irradiation. Sections were double-stained for apoptotic cells (red) and for F4/80 (blue) ($\times 164$). *a*, 1 h: numbers of apoptotic cells are increased relative to normal thymus (compare Fig. 2*a*); nearly all apoptotic cells are situated within F4/80⁺ macrophages; macrophages are slightly larger than in the normal thymus. *b*, 2 h: apoptotic cells are numerous; nearly all apoptotic cells are situated within enlarged macrophages. *c*, 4 h: as 2 h, except that some apoptotic cells are not inside macrophages. *d*, 24 h: apoptotic cells are rare; macrophages are smaller than at 2 and 4 h and are largely devoid of apoptotic bodies; the increased frequency of macrophages probably reflects shrinkage of the thymus. Relative to the normal thymus, cell yields from the irradiated thymuses were reduced by 5% at 1 h, 15% at 2 h, 30% at 4 h and 68% at 24 h (mean of 2 mice per group). All sections photographed at $\times 164$.

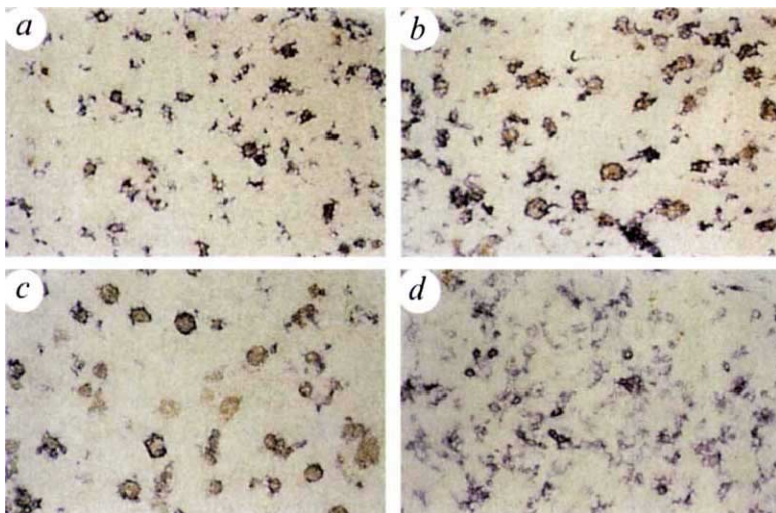
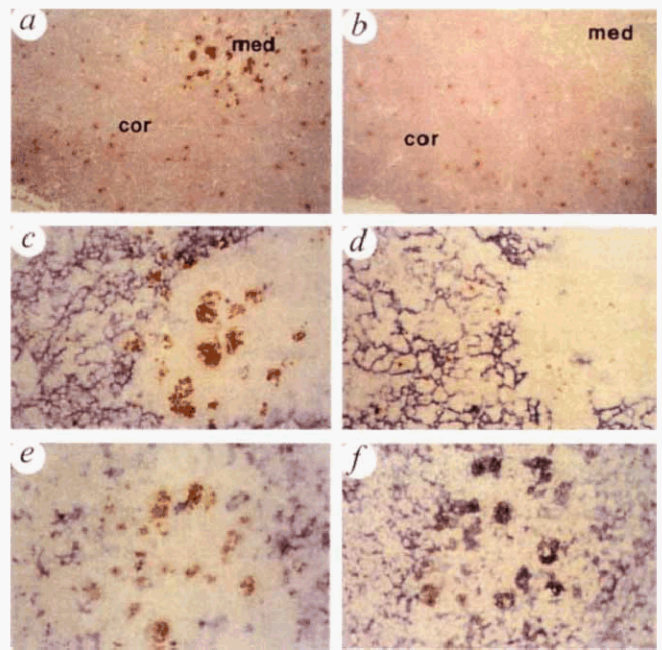


FIG. 4 Apoptotic cells in thymus of $I-E^+$ (a, c, e, f) compared with $I-E^-$ (b, d) V β 5 transgenic mice. a, $I-E^+$ thymus ($\times 88$): apoptotic cells show a normal distribution in the cortex (compare with Fig. 1) but are unusually prominent in the medulla; apoptotic cells in the medulla are much larger than in the cortex. b, $I-E^-$ thymus ($\times 88$): the distribution of apoptotic cells closely resembles the normal thymus (Fig. 1); apoptotic cells in the medulla are sparse. c, $I-E^+$ thymus ($\times 176$): double staining for apoptotic cells (red) and for 6C3 (blue), a marker for cortical epithelium, confirms that most of the enlarged apoptotic cells are confined to the medulla (which lacks 6C3 $^+$ epithelial cells) with some cells in the cortico-medullary junction. d, $I-E^-$ thymus ($\times 176$): double staining for apoptosis and 6C3 expression reveals almost no apoptotic cells in the 6C3 $^+$ medulla; some (small) apoptotic cells are visible within the network of epithelial cells in the 6C3 $^+$ cortex. e, $I-E^+$ thymus ($\times 176$): double staining for apoptotic cells (red) and for F4/80 (blue) reveals that the apoptotic cells in the medulla are not ingested by F4/80 $^+$ macrophages. f, $I-E^+$ thymus ($\times 176$): double staining for apoptotic cells (red) and for MAC-3 (blue), another macrophage marker, reveals that the apoptotic cells in the medulla are phagocytosed by MAC-3 $^+$ cells. METHODS. Cryostat sections of thymuses from TCR V β 5 transgenic mice⁹ bred to $I-E^-$ B6 mice or $I-E^+$ B6.107-1 mice²⁸ were single-stained for apoptotic cells as for Fig. 1, or double-stained for apoptotic cells and with F4/80, 6C3 or MAC-3 (ref. 29; Pharmingen) rat monoclonal antibodies as for Fig. 2.



involves apoptosis and clearance by a distinct population of F4/80 $^+$ MAC-3 $^+$ macrophages; why the macrophages that clear apoptotic cells in the cortex and medulla have different phenotypes is unknown. In spite of our findings on V β 5 transgenic mice, apoptotic cells in the medulla are rare in the normal thymus (Fig. 1b). This might reflect that physiological negative selection occurs primarily in the cortex rather than the medulla; if so, it is striking that the incidence of apoptotic cells in the cortex is not reduced in MHC-deficient mice. Whatever the explanation, T-cell death from negative selection seems to make only a small contribution to the background rate of apoptosis in the normal thymus. The vast majority of apoptotic cells in the thymus are probably a reflection of failure to undergo positive selection. □

Received 1 August; accepted 19 September 1994.

1. Robey, E. & Fowlkes, B. J. A. Rev. Immun. **12**, 675–705 (1994).
2. Shortman, K., Egerton, M., Spangrude, G. J. & Scollary, R. Semin. Immun. **2**, 3–12 (1990).
3. Tough, D. & Sprent, J. J. exp. Med. **179**, 1127–1135 (1994).
4. Scollary, R. & Shortman, K. in Recognition and Regulation in Cell-mediated Immunity (eds Watson, J. D. & Marbrook, J.) 3–20 (Dekker, New York, 1985).
5. Metcalf, D. in Ciba Symposium of the Thymus: Experimental and Clinical Studies (eds Wolstenholme, G. E. W. & Porter, P.) 242–263 (Churchill, London, 1966).
6. Kerr, J. F. R., Wyllie, A. H. & Currie, A. R. Br. J. Cancer **26**, 239–257 (1972).
7. Saint-Marie, G. & Peng, F. S. Rev. Can. Biol. **30**, 51–78 (1970).
8. Kendall, M. D. Immun. Today **5**, 286–287 (1984).
9. Fink, P. J., Swan, K., Turk, G., Moore, M. W. & Carbone, F. R. J. exp. Med. **176**, 1733–1738 (1992).
10. Woodland, D. L., Happ, M. P., Bill, J. & Palmer, E. Science **247**, 964–967 (1990).
11. Gavrieli, Y., Sherman, Y. & Ben-Sasson, S. A. J. Cell Biol. **119**, 493–501 (1992).
12. MacLennan, I. C. M. A. Rev. Immun. **12**, 117–139 (1994).
13. Cardell, S. C. et al. Adv. Immun. **55**, 423–440 (1994).
14. Savill, J., Fadok, V., Henson, P. & Haslett, C. Immun. Today **14**, 131–136 (1993).
15. Veis, D. J., Sorenson, C. M., Shutter, J. R. & Korsmeyer, S. J. Cell **75**, 229–240 (1993).
16. Duijvestijn, A. M. & Hoefsmit, E. C. M. Cell Tissue Res. **218**, 279–292 (1981).
17. Wyllie, A. H. Nature **284**, 555–556 (1980).
18. Anderson, R. E. & Warner, N. L. Adv. Immun. **24**, 215–335 (1976).
19. Smith, C. A., Williams, G. T., Kingston, R., Jenkinson, E. J. & Owen, J. J. T. Nature **337**, 181–184 (1989).
20. Shi, Y. et al. J. Immun. **146**, 3340–3346 (1991).
21. Murphy, K. M., Heimberger, A. B. & Loh, D. Y. Science **250**, 1720–1723 (1990).
22. Hengartner, H. et al. Nature **336**, 388–390 (1988).
23. Pullen, A. N., Kappler, J. W. & Marrack, P. Immun. Rev. **107**, 125–139 (1989).
24. Zijlstra, M. et al. Nature **344**, 742–746 (1990).
25. Grusby, M. J., Johnson, R. S., Papaioannou, V. E. & Glimcher, L. H. Science **20**, 1417–1420 (1991).
26. Surh, C. D. et al. J. exp. Med. **176**, 495–505 (1992).
27. Malek, T. R., Robb, R. J. & Shevach, E. M. Proc. natn. Acad. Sci. U.S.A. **80**, 5694–5698 (1983).
28. Murphy, D. B. et al. Nature **338**, 765–768 (1989).
29. Flotte, T. J., Springer, T. A. & Thorbecke, G. J. Am. J. Path. **111**, 112–124 (1983).

ACKNOWLEDGEMENTS. We thank D. Lo for TCR V β 5 and for 107-1 TG mice; L. Glimcher for I-A $^+$ mice; H. Kishimoto for help with FACS analysis; P. Le for introducing us to the TUNEL technique; B. Ernst and C. Reilly for breeding the mice; and B. Marchand for preparing the manuscript. This work was supported by grants from the US Public Health Service. C.D.S. is a Special Fellow of the Leukemia Society of America.

New member of the winged-helix protein family disrupted in mouse and rat nude mutations

Michael Nehls, Dietmar Pfeifer, Michael Schorpp, Hans Hedrich* & Thomas Boehm†

Molecular Medicine Group, Department of Medicine I, University of Freiburg, Hugstetter Strasse 55, D-79106 Freiburg, Germany

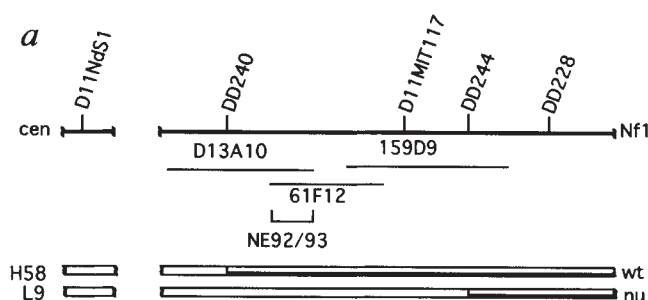
* Zentrale Tierversuchsanlage, Medizinische Hochschule Hannover, Konstanty-Gutschow-Strasse 8, D-30625 Hannover-Krefeld, Germany

MUTATIONS at the *nude* locus of mice and rats disrupt normal hair growth and thymus development^{1,2}, causing nude mice and rats to be immune-deficient. The mouse *nude* locus has been localized on chromosome 11 (refs 3, 4) within a region of <1 megabase⁵. Here we show that one of the genes from this critical region, designated *whn*, encodes a new member of the winged-helix domain family of transcription factors^{6,7}, and that it is disrupted on mouse *nu* and rat *rnuN* alleles. Mutant transcripts do not encode the characteristic DNA-binding domain, strongly suggesting that the *whn* gene is the *nude* gene. Mutations in winged-helix domain genes cause homeotic transformations in *Drosophila*⁸ and distort cell-fate decisions during vulval development in *Caenorhabditis elegans*⁹. The *whn* gene is thus the first member of this class of genes to be implicated in a specific developmental defect in vertebrates.

Previously we assigned the mouse *nude* locus to a region of less than 1 Mb within a yeast artificial chromosome (YAC) contig on mouse chromosome 11 (ref. 5). The derivation of a new microsatellite marker, DD244, allowed us to place the *nude* locus into an interval covered by only three YAC clones, D13A10, 61F12 and 159D9 (ref. 5; Fig. 1a). P1 recombinants spanning this region (M.N. et al., manuscript in preparation) were pooled and subcloned into λ GET, a novel phage-based exon-trap vector¹⁰. Several hundred clones from the resulting exon library

† To whom correspondence should be addressed.

FIG. 1 Chromosomal localization and structure of the mouse *whn* gene. **a**, The position of a trapped exon (designated NE92/93) within a YAC contig (see ref. 5 for details) spanning the mouse *nude* locus on chromosome 11. The genotype and phenotype of two informative F_2 mice from a (B6-nu $\times$ BALB/c) F_1 intercross⁵ is indicated at the bottom. Open bars show the *nude* chromosome, filled bars the wild-type allele. **b** (facing page), Composite cDNA sequence of the mouse *whn* gene. The exon boundaries are indicated by $\wedge$ underneath the last and first nucleotides of adjacent exons, respectively. The presumptive DNA-binding domain^{6,7} is underlined. An in-frame stop codon just upstream of the initiation codon¹⁸ is underlined; the termination codon in the *whn* gene is indicated by an asterisk. The single base-pair deletion detected in the mouse *nu* allele in exon 3 is indicated by a triangle; it is not possible to determine which G residue of the glycine codon (nucleotides 430–432) is deleted. This deletion causes a frameshift: the predicted aberrant protein terminates at a TGA stop codon (nucleotides 995–997) which is underlined. **c** (facing page), Comparison of protein sequences from the DNA-binding domains of several winged-helix domain genes; *Drosophila fkh* (ref. 8; 37% identity to *whn*), human HNF-3 γ (ref. 19) (31% identity to *whn*), human T-cell leukaemia virus enhancer factor (HTLF)¹⁵ (56% identity to *whn*), and the mouse *whn* gene. Differences from the *Drosophila fkh* sequence are noted; amino



acids identical in all four sequences are highlighted in the bottom line (con); gaps introduced to maximize alignment are indicated by dashes. METHODS. The (B6-nu $\times$ BALB/c) F_1 intercross has been described⁵. The DD244 marker detects a dinucleotide repeat ((GT)₂₄ in BALB/c mice) and is amplified as a 150-bp fragment using primers 5'-ACTCTG-GGAAGCCTCATGTC and 5'-CCTATGGTGCACAGCCACC. Other procedures have been described^{10,20,21}.

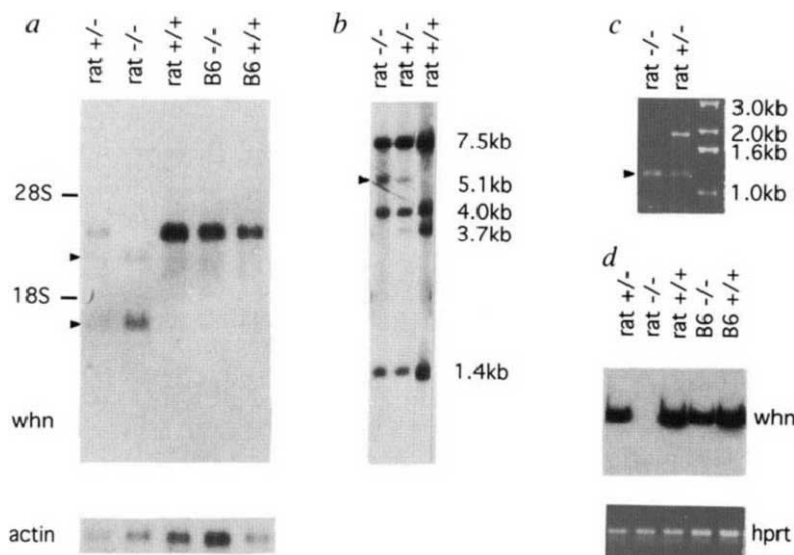
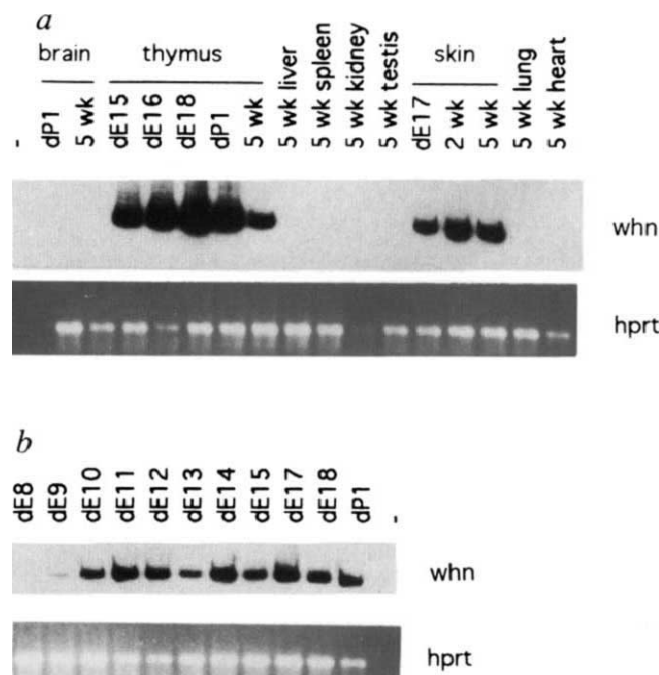


FIG. 2 Expression analysis of the *whn* gene. **a**, Expression of *whn* in individual tissues at various stages of development. RT-PCR¹⁰ was performed with primers derived from the 5' ends of exons 4 and 7, respectively. Note the presence of the 368-bp *whn* cDNA fragment only in thymus and skin. As a control, the RT-PCR result with *hprt*-specific primers is shown at the bottom. **b**, Time course of expression of *whn* during mouse embryonic development. Note that expression is detectable as early as dE9. RNA was isolated from the tissues indicated (**a**) or whole embryos (**b**) as described²⁰.

FIG. 3 Characterization of the rat *nuN* allele. **a**, Northern blot hybridization demonstrates the presence of aberrant *whn* transcripts in *nuN* animals. RNA was extracted from the skins of adult wild-type (+/+), nude (-/-) and heterozygous (+/-) Han: NZNU-*nuN* rats and wild-type (+/+) and nude (-/-) C57BL/6/J/Han (B6) mice, respectively. The two variant transcripts associated with the *nuN* allele are indicated by arrowheads. The probe used was a mouse *whn* cDNA spanning the entire coding region; as a control for loading and integrity of RNA, a mouse β -actin probe was used (bottom panel). **b**, Southern blot hybridization using the *whn* probe detects a rearranged *NcoI* fragment in the *nuN* allele. A 3.7-kb wild-type fragment is absent on the *nuN* allele and replaced by a 5.1-kb fragment (arrowhead); other restriction enzymes also revealed aberrations (data not shown), ruling out a simple restriction-fragment length polymorphism. **c**, Intragenic deletion on the *nuN* allele. Genomic DNA extracted from heterozygous (+/-) and nude (-/-) rats was amplified using primers derived from exons 4 and 5. Note the absence of the wild-type fragment in the nude rat; a fragment of unrelated sequence co-amplified as an internal control for the PCR reaction is indicated by an arrowhead. **d**, Absence of wild-type *whn* transcript in *nuN* rats. RT-PCR was performed with primers derived from the 5'-ends of exons 4 and 7, respectively. The 368-bp *whn* cDNA fragment is absent in nude rats. As a control, the RT-PCR result with *hprt*-specific primers is shown at the bottom. RNAs were isolated from skin tissue²⁰; about 0.5 μ g of poly(A)⁺-selected RNA was used per lane. Probe labelling, hybridization and PCR were done as described^{10,20}.



b whn CAGACGGTGGAGCTCCTGGCCCCCAGACCCAGGCCCCACGCCGACCTGCTTCACTGAGCAGCTGGCTTTCTTCGAGGCCA 83
 exon 1
 M V S L L P P Q S D V T L P G S T R L E G E P Q G D L
 GGACTGGGTGATGGTGTGCTACTCCTCCGCGAGTCTGACGTACACCTTCAGGCTCCACCCGACTGGAGGGCGAACCCTAAGGGGACCT 173
 exon 2
 M Q A P G L P D S P A P Q N K H A N F S C S S F V P D G P P
 CATGCAGGCTCCGGGCTCCAGACTCCCTGCCCCACAGAACAAGCATGCTAACTTCAGCTGCTCGTCTGTTGTGCTGACGGCCCTCC 263
 E R T P S L P P H S P S I A S P D P E Q I Q G H C T A G P G
 AGAGAGGACACCCTCACTGCCCCACAGACCCAGCATCGCATCTCCAGACCCAGAGCAGATCCAGGGCCACTGCACAGCCGGACCCGG 353
 P G S F R L S P S E K Y P G F G F E E G P A G S P G R F L K
 CCCGGGCTCCTTCCGCCCTTTCTCCTTCAGAAAAGTATCTGGCTTTGGCTTTGAGGAGGGCCAGCAGGCAGCCAGGGCGCTTCTCTAA 443
 exon 3
 G N H M P F H P Y K R H F H E D I F S E A Q T A M A L D G H
 GGGCAACCACATGCCTTTCCACCCCTACAAGAGGCACCTCCATGAGGACATCTTCTCTGAGGCCAGACGGCCATGGCACTTGATGGACA 533
 S F K T Q G A L E A F E E I P V D M G D A E A F L P S F P A
 CTCCTTTAAGACTCAGGGGGCACTGGAAGCCTTTGAGGAGATCCCTGTGGACATGGCGATGCTGAGGCCCTTCTGCTTAGCTTCCCAGC 623
 E A W C N K L P Y P S Q E H N Q I L Q G S E V K V K P Q A L
 AGAGGCTTGGTGCAATAAATCCCTTACCCAGCCAGGAACACAACCAATTCAGAGGGTCAAGGTCAAGGCCCAAGCTCT 713
 D S G P G M Y C Y Q P P L Q H M Y C S S Q P A F H Q Y S P G
 GGACAGTGGTCTGGGATGTACTGCTACCAGCTCCCTTGCAACATATGACTGTCTTCTCAGCCTGCCTTCCATCAGTACTCCCCGGG 803
 exon 4
 G G S Y P V P Y L G S P H Y P Y Q R I A P Q A N A E G H Q P
 TGGAGGCAGCTACCTGTGCCCTACCTGGGCTCACCTCACTATCCCTATCAGAGGATTGCACCCAGGCCAACGCCGAAGGTCAACAGCC 893
 exon 5
 L F P K P I Y S Y S I L I F M A L K N S K T G S L P V S E I
 ACTCTTCCAAAGCCATCTACTCTTACAGCATCCTCATCTTTCATGGCCCTTAAGAACAGTAAGACCGGAAGCCTTCCAGTCACTGAAAT 983
 Y N F M T E H F P Y F K T A P D G W K N S V R H N L S L N K
 CTACAATTTCACTGAGCGAGCACTTCCCTTACTTCAAGACTGCTCCTGATGGCTGGAAGAATCTGTTCGCCATAACCTGTCCCTCAACAA 1073
 C F E K V E N K S G S S S R K G C L W A L N P S K I D K M Q
 GTGCTTTGAGAAGGTGGAGAATAAATCCGGAAGTTCCTCTCGAAAGGGCTGTCTGTGGGCCCTCAATCCTTCCAAAATCGACAAGATGCA 1163
 exon 7
 E E L O K W K R K D P I A V R K S M A K P E E L D S L I G D
 GGAAGAACTGCAGAAGTGAAGAGGAAGACCCCATTTGCTGTGCGCAAAAGCATGGCCAAACCAGAAGAGCTGGACAGCCTCAITGGAGA 1253
 K R E K L G S P L L G C P P P G L A G P G P I R P M A P S A
 CAAAAGGGAAAACTGGGCTCTCCGCTGCTGGGCTGTCCACCCCTGGGCTGGCAGGCCAGGTCCCATCCGGCCCATGGCACCATCAGC 1343
 G L S Q P L H P M H P A P G P M P G K N P L Q D L L G G H A
 TGGTCTTTCCAGCCTCTGCACCCATGCATCCAGCTCCAGGCCCATGCTGGCAAGAACCCCTGCAGGACCTACTGGGTGGCCATGC 1433
 exon 8
 P S C Y G Q T Y P H L S P S L A P S G H Q Q P L F P Q P D G
 TCCCTCTGCTATGGGCAGACCTACCCACACCTTTCCCCCAGCTGGGCCCTTCTGGACACAGCAGCCATTGTTCACAGCCAGATGG 1523
 H L E L Q A Q P G T P Q D S P L P A H T P P S H G A K L M A
 GCATCTTGAGCTGCAGGCCAGCCAGGCACCCCCAGGACTCACCTCTACCTGCCCACACACCACCCAGCCACGGTGGCAAGCTGATGGC 1613
 E P S S A R T M H D T L L P D G D L G T D L D A I N P S L T
 TGAGCCTTCTCAGCCAGGACCATGCACGATACTCTACTGCCAGATGGAGACCTTGGGACTGACCTGGATGCTATCAACCTTCTCTAC 1703
 D F D F Q G N L W E Q L K D D S L A L D P L V L V T S S P T
 TGACTTCGACTTCAGGGAATCTGTGGGAGCAGCTGAAGGATGACAGCTTGGCCCTGGACCCCTCGTATTGGTGACCTCGTCCGCCGAC 1793
 S S S M L P P P P A A H C F P P G P C L A E T G N E A G E L
 GTCATCTCCATGTTGCCACCCCCACAGCAGCCATTGCTTCCCCCAGGGCCTTGTCTGGCAGAAACAGGCAATGAGGCAGGTGAAGT 1883
 exon 9
 A P P G S G G S G A L G D M H L S T L Y S A F V E L E S T P
 GGCACCTCCAGGCAGCGGGCTCCGGTGCTTGGGAGACATGCACCTCAGCACTCTCTACTCCGCCCTTGTGGAAGTGGAGTCCACGCC 1973
 S S A A A G P A V Y L S P G S K P L A L A *
 CTCCTCAGCAGCTGCCGGCCCTGCCGTGTACCTCAGTCCCGGCTCAAAGCCATTGGCTCTGGCTTGAGCTATGCCAACATCAGCTGCGG 2063
 CCTTTGCCCTAGCTGGCTGCCCATATTGGGCTCACCTTAAAGGTCAAAGAAGGAAACACTACCTGTCTCCTATGCCACTCAGCCAACTT 2153
 ATTTGTTAGCCAGAAGTTGGGGGCATCCATCTAGGATGCTACCTGGTGGACCCGCATCTCACACGGGGCCCCCAGCAAGGAAAGTGTGCG 2243
 AAAAGAAGCAACAGCGCGCCCTTAGCGCCAGCTCAGTGAAGTTCAGCTCTCGAGCGTGAAGTCAAGCTCTACACACTGCTCATCATG 2333
 TGCCTTACCTCAGAGGAAGCTTGGGTACAGAGATCTGATTTGATTTCTGGGGCAGCTGAGAGCTAAAAGCTTTAGTTAGCAAAAGCTCA 2423
 GGGCAGTGTGACAGGTCAAAGATCACCCCTCAAACCTCATCCAATCCCCATGTGCTCTACAGACTTGGCCCTCTCCC 2503

C KPPYSYISLITMAIQNPNTRMLTLSEIYQFIMDLFPFYRQNRQNSIRHSLSFNDCFVKIPRTDPKP-GKGSFWTLHPDSGNMFENG CYLRR fkh
 QA GK W Y E VA S - Y A S HNF
 FSL Y EHS NKC PVK SW L H YFATAPTG K V N LNK Q VE SHG VN L CVD EYKPNLI-QA- KK HTLF
 I SI F LK SK GS PV NFMTEH YFKTAPDG K V N LNK E VENKSGSSSR CL ALN SKIDKMQ-EE- QK whn
 KP YS LI MA L EIY FP W NS RH LS N CF K KG W L con

FIG. 4 Structure of aberrant *rnuN whn* transcripts. The relevant parts of cDNA sequences from wild-type rat *whn*, and two variant transcripts (X and Y) emanating from the *rnuN* allele are shown. All cDNAs are of wild-type sequence up to the end of exon 4. The wild-type transcript then continues into exon 5, whereas the variant transcripts show completely unrelated sequences. Both are unrelated to the first 200 base-pairs of sequence in the intron following exon 4. The predicted protein sequences (mutant forms, lower-case) are indicated above the nucleotide sequences. Note the presence of only a few conservative amino-acid changes (underlined) in the wild-type rat sequences compared with the mouse *whn* gene. The predicted translational stops in the variant cDNAs are marked asterisks.

METHODS. A cDNA library from poly(A)⁺ mRNA isolated from the skin of heterozygous Han: NZNU-*rnuN* rats was prepared in λ gt10 and screened with a full-length mouse *whn* cDNA²⁰.

```

.. E H N Q V L O G S E V K V K P Q A L D N G P
.. GAACACAACCAAGTCTTCAGGGGTGAGGTCAGGTCAGGCCCAAGCTTTGACAAATGGTCT
  rat exon 3      ^^

G M Y C Y Q P P L Q H V Y C S S Q P T F H Q
GGGATGTACTGCTACAGCCTCCCTTGACAGCATGTGTACTGTTCTTCAGCCACCTTTCATCAG
  rat exon 4      ^

wild-type rat exon 5      Y S P G G G S Y P V P Y L G ..
                        TACTCCCCGGGTGGAGGAGTACCTGTGCTTACCTACCTGGGC..

a k r p t g d t e r n l l t a d s r l c p g d p *
x GCTAAAAGACCTACAGGTGATACAGAAAGAAATCTGCTACAGCTGACAGCAGGCTATGCCCGGAGACCCCTGA..

p l c p t e s i p r n r t a q a *
y CCATTGTGTCTACAGAGCATTCAGAAACCGACAGCCCAAGCCTTA..

```

were sequenced and localized to one of about 40 intervals defined by clone ends in the YAC and P1 contig (M.N. *et al.*, manuscript in preparation). To identify exons from the *nude* gene, the expression pattern of physically ordered exons was analysed by reverse transcription and polymerase chain reaction (RT-PCR). We reasoned that the putative *nude* gene should be preferentially, if not exclusively, expressed in the thymic anlage (and later in thymus) and the skin, the sites of phenotypic abnormalities in nude animals^{11,12}. Of more than 50 exons analysed in this way, only one (designated NE92/93) showed specific expression in thymus and skin, with no detectable activity in several other organs, including brain, liver and spleen (Fig. 2a). This exon was positioned to the overlap between YACs D13A10 and 61F12 (Fig. 1a), within the interval containing the *nude* gene, and was thus a good candidate for the *nude* locus. Because the thymic rudiment is defective in nude animals¹³, it is not colonized by T-cell progenitors. The expression of the NE92/93 transcription unit is already detectable in the mouse at embryonic day 9 (dE9) (Fig. 2b), which is before T-cell progenitors have migrated to the thymic rudiment (dE10/11; ref. 14). This timing of expression is expected for the *nude* gene, whose function is essential for differentiation of thymic epithelium¹¹.

To characterize further the transcription unit containing the NE92/93 exon, complementary DNA clones were isolated from a mouse skin cDNA library, and the genomic structure of this gene was determined (Fig. 1b). Sequence analysis showed that exon NE92/93 represents exon 4 of this gene (Fig. 1b). The assembled cDNAs predict a protein of 648 amino acids with homology to the winged-helix domain family of transcription factors^{6,7} (Fig. 1c). The presumptive DNA-binding domain (underlined in Fig. 1b) has the highest similarity to T-cell leukaemia virus enhancer factor (56% identities), a human winged-helix domain protein¹⁵. The protein sequences C-terminal to the DNA-binding domain are characterized by a large number of negatively charged amino acids and a high proportion of proline residues. These features are reminiscent of transcriptional activation domains¹⁶ and compatible with the role of winged-helix domain proteins as transcriptional regulators^{6,7}.

To investigate the possible presence of mutations in this gene, designated *whn* (for winged-helix nude), we first studied C57BL/6/J/Han (abbreviated B6) congenic mice carrying the original *nude* allele⁵. To this end, all eight coding exons and the flanking intron sequences were sequenced in B6 wild-type and nude animals. The only difference observed is a single base-pair (G) deletion in exon 3 of the *whn* gene (Fig. 1b). This mutation causes a frameshift, and the predicted aberrant protein terminates at a TGA stop codon in exon 6 (Fig. 1b) without encoding the DNA-binding domain. This finding indicated that the *whn* gene might indeed be the *nude* gene.

To test this hypothesis, we analysed a *nude* allele in the Han: NZNU-*rnuN* rat², because nude rats have an identical phenotype to nude mice. Furthermore, the *nude* locus in rats has been

mapped to chromosome 10, in a region homologous to mouse chromosome 11 (ref. 17). The fact that the rat and mouse *whn* genes are almost identical (data not shown) facilitated this analysis. In contrast to the mouse *nu* allele, the *rnuN* allele is characterized by the presence of aberrant transcripts from the *whn* gene (Fig. 3a). The normal transcript of about 4 kilobases (kb) is replaced by two variant transcripts of about 3 kb and 1.5 kb, respectively. Analysis of genomic DNA from nude and heterozygous rats suggested that there was an intragenic deletion encompassing exon 5 (Fig. 3b, c) and exon 6 (data not shown), which encodes the N-terminal half of the DNA-binding domain (Fig. 1b). In addition, RT-PCR using primers located in exons 4 and 7, respectively, did not yield a detectable product in RNA isolated from the skin of nude animals (Fig. 3d), suggesting aberrant splicing on the *rnuN* allele. Sequence analysis of the two variant transcripts from the *rnuN* allele indicated that, as expected, the mutant transcripts diverge from the wild-type sequence after exon 4, namely before the exon sequences encoding the presumptive DNA-binding domain (Fig. 4). The appended sequences, presumably derived from sequences harbouring cryptic splice acceptor sites, predict translational stops after 24 and 12 amino acids, respectively. Therefore, no functional *whn* protein can be made from *rnuN* transcripts.

The mapping and sequencing results presented here provide strong genetic evidence that the nude phenotype in the mouse is the result of a deleterious mutation in the *whn* gene. This conclusion is supported by the temporal and spatial expression pattern of *whn*. We also demonstrate that this gene is disrupted in a rat strain with an identical phenotype. Neither of the two *nude* alleles studied can give rise to a protein containing the presumptive DNA-binding domain of the *whn* protein, which is reminiscent of the homeotic loss-of-function mutations in the *Drosophila* fork head (*fkh*) gene⁸. Although our results give no indication of the pathophysiology of the nude phenotype, they suggest that the *whn* protein regulates tissue-specific transcription, and/or cell fate in the thymic anlage and the hair papilla by activating downstream target genes. □

Received 14 June; accepted 25 August 1994.

1. Lyon, M. F. & Searle, A. G. *Genetic Variants and Strains of the Laboratory Mouse* (Oxford Univ. Press, Oxford, UK, 1989).
2. Hedrich, H. (ed.). *Genetic Monitoring of Inbred Strains of Rats* (Fischer, Stuttgart, Germany, 1990).
3. Takahashi, Y. *et al.* *J. exp. Med.* **175**, 873–876 (1992).
4. Lisitsyn, A. N. *et al.* *Nature Genet.* **6**, 57–63 (1994).
5. Nehls, M. *et al.* *Eur. J. Immun.* **24**, 1721–1723 (1994).
6. Weigel, D. & Jäckle, H. *Cell* **63**, 455–456 (1990).
7. Lai, E., Clark, K. L., Burley, S. K. & Darnell, J. E. Jr *Proc. natn. Acad. Sci. U.S.A.* **90**, 10421–10423 (1993).
8. Weigel, D., Jürgens, G., Küttner, F., Seifer, E. & Jäckle, H. *Cell* **57**, 645–658 (1989).
9. Miller, L. M., Gallegos, M. E., Morisseau, B. A. & Kim, S. K. *Genes Dev.* **7**, 933–947 (1993).
10. Nehls, M., Pfeifer, D. & Boehm, T. *Oncogene* **9**, 2169–2176 (1994).
11. Cordier, A. C. & Haumont, S. M. *Am. J. Anat.* **157**, 227–263 (1980).
12. Köpf-Maier, P., Mbonenko, V. F. & Merker, H. J. *Acta anat.* **139**, 178–190 (1990).
13. van Vliet, E. *et al.* *Eur. J. Immun.* **15**, 675–681 (1985).
14. Palacios, R. & Samaridis, J. *Eur. J. Immun.* **21**, 109–113 (1991).

15. Li, C., Lusis, A. J., Sparkes, T., Tran, S. M. & Gaynor, R. *Genomics* **13**, 658–664 (1992).
16. Mitchell, P. J. & Tijan, R. *Science* **245**, 371–378 (1989).
17. Cash, J. M. et al. *Mammal. Genome* **4**, 37–42 (1993).
18. Kozak, M. *Nucleic Acids Res.* **15**, 8125–8148 (1987).
19. Lai, E., Prezioso, V. R., Tao, W., Chen, W. S. & Darnell, J. E. Jr *Genes Dev.* **5**, 416–427 (1991).
20. Boehm, T. et al. *Proc. natn. Acad. Sci. U.S.A.* **88**, 4367–4371 (1991).
21. Sanger, F., Nicklen, S. & Coulson, A. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463–5467 (1977).

ACKNOWLEDGEMENTS. We thank K. Lüne, S. Krause, U. Matysiak-Schölze, H. Dierbach, C.-M. Prokop and G. Igloi for their help. Supported by a grant from the Deutsche Forschungsgemeinschaft. The mouse *whn* cDNA sequence has been deposited with EMBL/GenBank/DBJ data-bases (accession number X81593).

A new class of retinoids with selective inhibition of AP-1 inhibits proliferation

Andrea Fanjul, Marcia I. Dawson*†, Peter D. Hobbs*, Ling Jong*, James F. Cameron*, Eli Harlev*, Gerhart Graupner, Xian-Ping Lu & Magnus Pfahl†

Cancer Center, La Jolla Cancer Research Foundation, La Jolla, California 92037, USA

* Life Sciences Division, SRI International, Menlo Park, California 94025, USA

RETINOIDS regulate many biological processes, including differentiation, morphogenesis and cell proliferation^{1–3}. They are also important therapeutic agents, but their clinical usefulness is limited because of side effects^{4–8}. Retinoid activities are mediated by specific nuclear receptors, the RARs and RXRs, which can induce transcriptional activation through specific DNA sites^{3,9–11} or by inhibiting the transcription factor AP-1 (refs 12–15), which usually mediates cell proliferation signals¹⁶. Because the two types of receptor actions are mechanistically distinct^{12,15}, we investigated whether conformationally restricted retinoids, selective for each type of receptor action, could be identified. Here we describe a new class of retinoids that selectively inhibits AP-1 activity but does not activate transcription. These retinoids do not induce differentiation in F9 cells but inhibit effectively the proliferation of several tumour cell lines, and could thus serve as candidates for new retinoid therapeutic agents with reduced side effects.

Conformationally restricted synthetic retinoids that selectively enhance transcriptional activation by retinoic acid receptor (RAR) subtypes^{17–19} or retinoid X receptor (RXR) homodimers²⁰ are assumed to recognize differences in the ligand-binding pockets of these receptors, whereas the flexible polyene side chains of the natural retinoids all-*trans*-RA (*t*-RA) and 9-*cis*-RA allow fairly indiscriminate receptor interactions^{17–21}. We examined whether conformationally restricted retinoids could also differentially activate receptor transcriptional activation or anti-AP-1 activity.

We compared the transcriptional activation activity of about 50 synthetic retinoids with their anti-AP-1 activity using RAR α , β , γ and RXR α (Fig. 1a). Although *t*-RA and several synthetic retinoids (not shown) behaved very similarly in transcriptional activation and anti-AP-1 assays (Fig. 1a), some retinoids clearly demonstrated distinct receptor selectivities in both assays (Fig. 1b). For instance, SR11105 is RAR γ , β -selective in transactivation¹⁷ but RAR α , β -selective in the anti-AP-1 assay, and SR11217 is a strong RXR α -selective compound in transactivation²⁰ but is RAR β -selective in the anti-AP-1 assay (Fig. 1b). These results demonstrate that retinoids which were can display distinct receptor selectivities in the two different receptor-mediated responses.

We next investigated a series of retinoids which were inactive in the transcriptional activation assay. Interestingly, several of these 'inactive' compounds turned out to be potent activators of the RARs or RXR α in the anti-AP-1 assay. As observed for *t*-RA and 9-*cis*-RA, the activity of the retinoids was dependent on the cotransfected receptors. The most potent anti-AP-1-selective compounds in this series and their receptor specificities are compiled in Table 1. Although these retinoids show anti-AP-1 activity with more than one of the receptors, a certain degree of receptor selectivity was usually observed (Fig. 1c and Table 1). When these retinoids were examined in C₃H/T101/2 (mouse embryo fibroblast) cells, essentially the same results (not shown) as with HeLa cells were obtained.

We tried to find retinoids with no anti-AP-1 activity, but which are potent as transcriptional activators. From a large number of transcriptionally active retinoids, we identified one compound, SR11235, which induced essentially no (less than 20%) receptor anti-AP-1 activity. As previously shown²⁰ SR11235 is an RXR α -selective transcriptional activator (Fig. 1d and Table 1). Thus retinoids selective either for anti-AP-1 activity or transcriptional activation can be identified.

Retinoids that activate only anti-AP-1 activity are likely to interact with at least part of the ligand-binding pocket of the receptor. Such retinoids could therefore be expected to have antagonist activity in the transcriptional activation assay. This was indeed the case for SR11302, which effectively inhibited *t*-RA-induced transcriptional activation by RAR α and RAR γ but not by RAR β , consistent with the observation that this retinoid showed strong anti-AP-1 activity with RAR α and RAR γ but not with RAR β (Fig. 1e). A similar correlation between antagonism of transcriptional activation and anti-AP-1 activity was obtained for several other compounds (not shown). These data support the assumption that the same ligand-binding pocket is likely to mediate transcriptional activation and anti-AP-1 activity. However, the anti-AP-1 selective retinoids were only weak antagonists compared to other antagonists^{22,23}.

The biological activities of anti-AP-1-selective retinoids were examined in the embryonal carcinoma cell line F9, in which RA is known to induce differentiation, and in cancer cell lines whose growth can be inhibited by retinoids. As judged from cell proliferation assays and morphological changes, the anti-AP-1-selective retinoids SR11220, SR11238 and SR11302 failed to inhibit the proliferation of F9 cells (Fig. 2a) or to induce their differentiation (Fig. 2b,d–g) as observed with *t*-RA and 9-*cis*-RA (Fig. 2b,a, b). In contrast, these same retinoids were potent inhibitors of the growth of the lung cancer cell line Calu-6 and the breast cancer cell line T-47D (Fig. 3). HeLa cells were also inhibited, although growth inhibition of these cells by *t*-RA and the synthetic retinoids was weaker (Fig. 3). SR11235, the transcriptional activation-selective compound, did not inhibit either one of these cell lines. The anti-AP-1-selective compounds were often more potent inhibitors of these cancer cell lines than *t*-RA, especially at lower concentrations. In addition, the activity of the individual compounds varied somewhat from cell line to cell line, consistent with their different receptor contents (ref. 24 and our unpublished data). Thus the anti-AP-1 activities of SR11220, SR11238 and SR11302 can be correlated with their anti-proliferative activity. The observation that these same retinoids do not induce differentiation in F9 cells is consistent with the fact that RA-induced differentiation is associated with the upregulation of several genes through retinoic acid receptor enhancers (RAREs)^{25,26}.

We have shown that retinoids that can induce selectively only one of the two major receptor activities, transcriptional activation or anti-AP-1 activity, can be identified. The two distinct mechanisms can thus be expected to require different receptor configurations induced or stabilized by the ligand. *t*-RA with its flexible polyene side chain is apparently able to accommodate both receptor configurations, whereas the conformationally restricted retinoids identified here favour either the anti-AP-1

† To whom correspondence should be addressed.

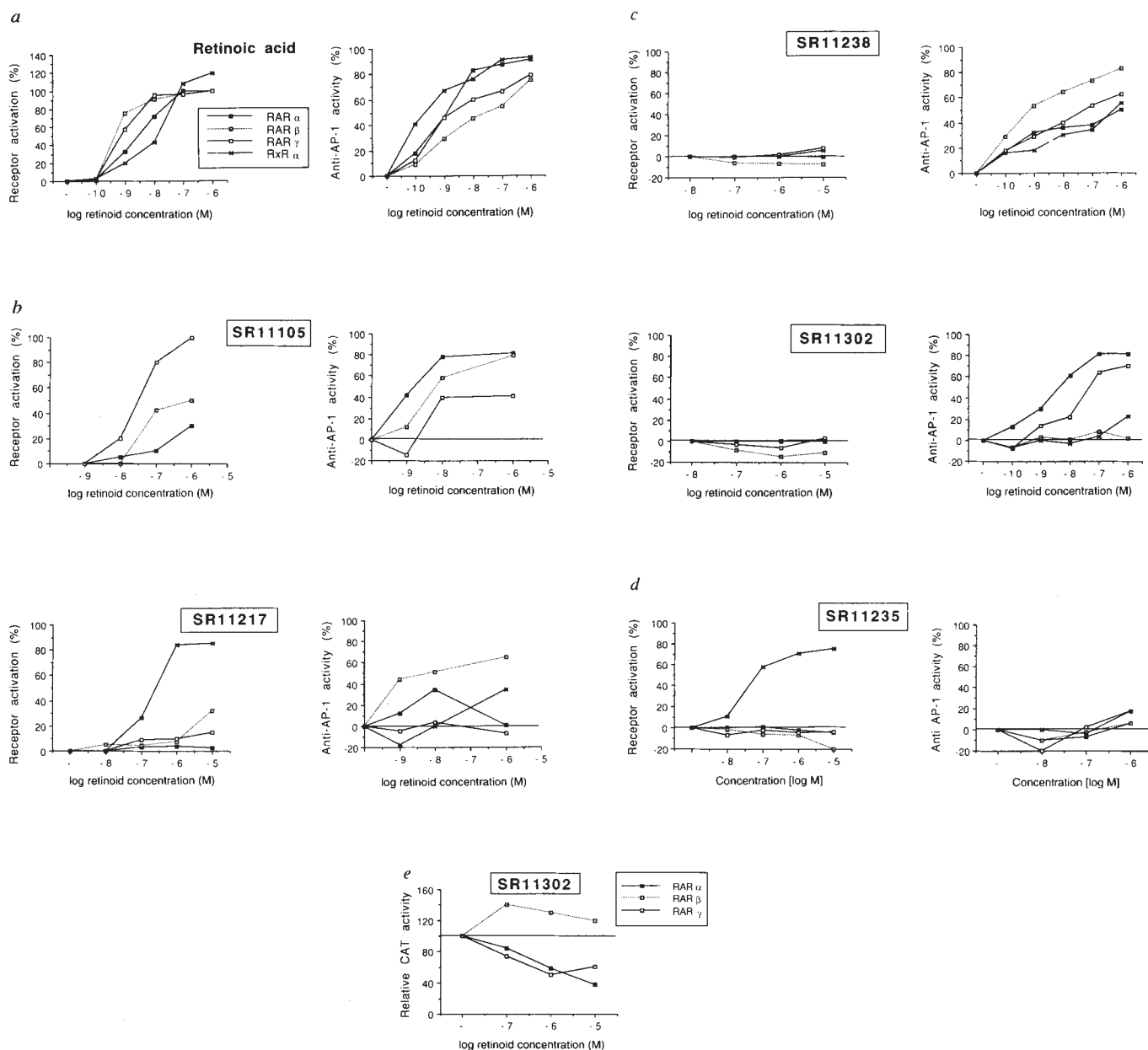


FIG. 1 Transcriptional activation and anti-AP-1 activities of retinoids. **a-d**, Transient transfection assays were used to determine transcriptional activation and anti-AP-1 activities of various retinoids. In addition to those retinoids SR11235, SR11238 and SR11302 named in Table 1, SR11105 (E)-4-(2-(5,5-dimethyl-5,6,7,8-tetrahydro-3-naphthalenyl)-propenyl)benzoic acid; and SR11217, 4-(2-methyl-1-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)propenyl)benzoic acid were evaluated. Transcriptional activation and repression are expressed as relative activation, with 100% representing the activity observed with RARs in the presence of 10^{-6} M *t*-RA, and 100% defining the RXR activity observed in the presence of 10^{-6} M 9-*cis*-RA. Under the conditions used here *t*-RA at 10^{-6} M increased RAR α activity 25-fold, RAR β and RAR γ were somewhat less induced because of their constitutive basal activity¹⁶. Induction by RXR α in the presence of 10^{-6} M 9-*cis*-RA was 13-fold. TPA (100 ng ml⁻¹) resulted in an 18-fold induction of the Col-CAT reporter. **e**, Anti-AP-1-selective retinoids show antagonist activity. Transient transfections into CV-1 cells were used to measure antagonist activity of anti-AP-1-selective retinoids. After transfection, cells were grown in medium containing 10^{-9} M *t*-RA alone or *t*-RA and the indicated concentrations of the synthetic retinoids. 100% CAT activity represents the activity measured in the presence of *t*-RA only. Data

represent results from three separate experiments done in duplicate. **METHODS.** To measure receptor activation, CV-1 cells were transfected with 50 ng of expression plasmids for the receptors RAR α , β , γ and RXR α , 100 ng of the reporter gene (TREpal)₂-tk-CAT and 150 ng of β -galactosidase expression plasmid pCH110. The TREpal is a RARE that allows for optimal transcriptional activation by all three RARs as well as RXR α . Cells were grown for 24 h in 10% charcoal-treated DME medium containing various concentrations of the retinoids to be analysed. In general, the assays were done as described²⁸. Receptor activation was also analysed in HeLa cells using the ApoA1-RARE-CAT reporter²⁰, yielding essentially the same results as obtained with the TRE reporter in CV-1 cells. To determine anti-AP-1 activity of retinoids, HeLa cells were transiently transfected with 100 ng of -73ColCAT (containing an AP-1 site) reporter gene¹² and 50 ng of expression plasmids for RAR α , β , γ and RXR α . After transfection, cells were grown in 0.5% charcoal-treated fetal calf serum in the presence or absence of retinoids with or without TPA (100 ng ml⁻¹) for 24 h before collecting. The β -galactosidase expression vector (pCH110, Pharmacia) was cotransfected to normalize for CAT activity. The assays were as described previously^{12,14}.

TABLE 1 Structures and activities of retinoids

Retinoid		Anti-AP-1 activity (%) at 10^{-6} M	Anti-AP-1 selectivity at 10^{-6} M	Transactivation activity (%) at 10^{-5} M
Structure	Name			
	<i>trans</i> -RA		RAR α , β , γ	
		100		100
	9- <i>cis</i> -RA		RXRs	
	SR11238		RAR β > RAR γ > RXR α $\geq$ RAR α	10 (RAR γ , RXR α)
		> 81		
	SR11302		RAR α > RAR γ > RXR α = RAR β	0
	SR11220		RAR α > RAR γ > RAR β > RXR α	< 20 (RAR γ , RAR α)
	SR11327	80–71	RAR α > RAR β > RAR γ > RXR α	< 15
	SR11228	70–61	RAR γ > RAR α > RXR α > RAR β	< 15 (RAR β , RAR γ)
	SR11324	60–50	RXR α > RAR γ > RAR β > RAR α	< 20 (RAR γ , RAR β)
	SR11235	< 22	RAR γ = RXR α > RAR β = RAR α	75 (RXR α)

Transcriptional activation and anti-AP-1 activity were evaluated as shown in Fig. 1.

* The numbers give the activity for the receptor subtype (see adjacent column) with the highest activity. SR11238, 2-(3,4-dihydro-4,4-dimethyl-2H-1-benzopyran-6-yl)-2-(4-carboxyphenyl)-1,3-dithiane; SR11302, (2E,4E,6Z,8E)-3-methyl-7-(4-methylphenyl)-9-(2,6,6-trimethylcyclohexen-1-yl)-2,4,6,8-nonatetraenoic acid; SR11220, methyl (Z)-4-(1-acetoxy-2-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)ethenyl)benzoate; SR11327, (E)-4-(2-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-3-phenylpropenyl)benzoic acid; SR11228, 5-((5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)carbonyl)-2-naphthalenecarboxylic acid; SR11324, 4-(3-(4-dimethylaminophenyl)methyl-5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-anthracenyl)benzoic acid, hydrochloride salt; SR11235, 2-(5,6,7,8-tetrahydro-5,5,8,8-tetramethyl-2-naphthalenyl)-2-(4-carboxyphenyl)-1,3-oxathiolane.

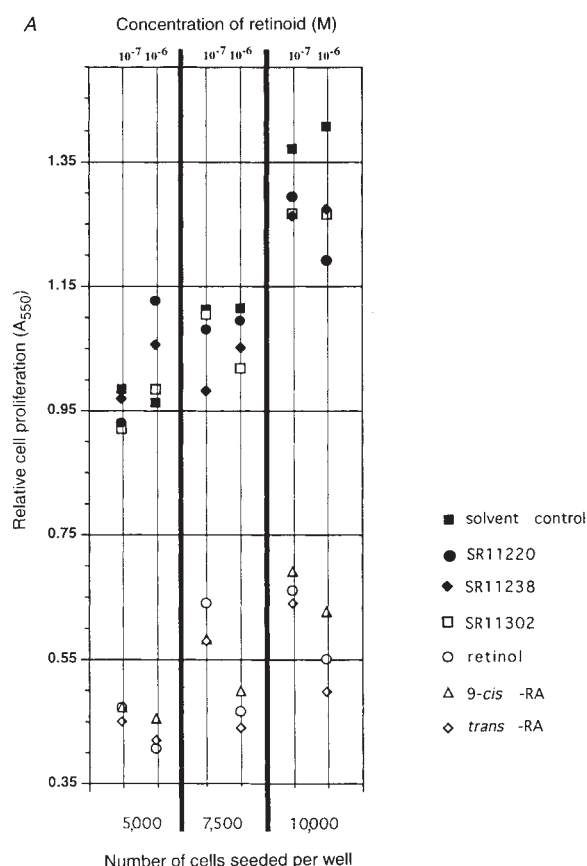
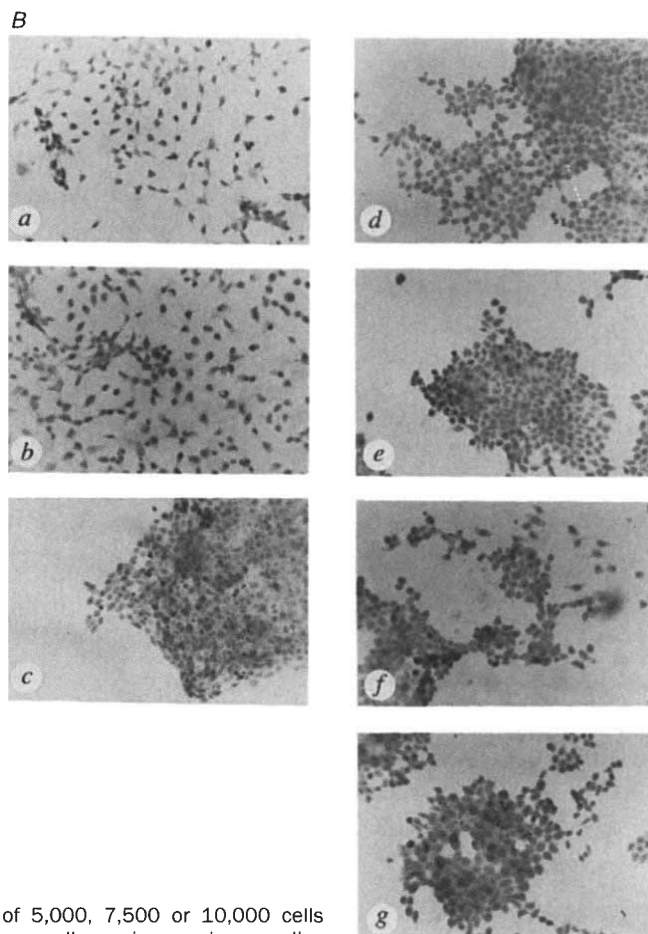


FIG. 2 Biological activities of anti-AP-1-selective retinoids. A, F9 cell proliferation. F9 cells were seeded in plastic wells (24-well plates, Costar) coated with 0.1% gelatin at different densities as indicated. After adhesion (day 0), cells were treated (day 1) with 10^{-7} M or 10^{-6} M retinoid. Solvent controls received an identical dilution of ethanol only (0.01% or 0.001% final concentration of ethanol). Treatment was repeated 48 h after the initial treatment (day 3). Cells were assayed on day 6 with a commercial cell proliferation kit (Promega) for their capability to reduce MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) dye to a coloured formazan product as an index of cell survival and proliferation. Absorption at 550 nm was determined and plotted against retinoid concentration at the respective cells densities. B, Morphological changes in F9 cells. Cells were seeded at densities



of 5,000, 7,500 or 10,000 cells per well on microscopic coverslips coated with 0.1% gelatin and treated with retinoids as in A. On day 6, cells were fixed in acetone at -20°C for 5 min, stained with 0.1% trypan blue (Sigma) in 10% acetic acid/methanol at room temperature for 2 h, and mounted. Photomicrographs were taken through a $\times 25$ objective (total magnification $\times 62.5$) at tungsten lamp illumination and recorded on Kodak B&W 400 ASA film. Shown are data (a-g) from treatment with 10^{-6} M retinoid. a, All-trans-retinoic acid; b, 9-cis-retinoic acid; c, solvent control; d, SR11220; e, SR11327; f, SR11238; g, SR11302. Essentially identical results were seen after treatment with 10^{-7} M retinoid.

configuration (SR11238, SR11302 and so on) or the transcriptional activation configuration (SR11235) of the receptor.

Not unexpectedly, anti-AP-1-selective retinoids have a reduced range of biological activities and cannot induce differentiation in F-9 cells, although they have maintained the ability to inhibit efficiently proliferation of cancer cells. In fact they appear to be more potent inhibitors of proliferation than *t*-RA in several cell lines. The observed inhibition is specific and not due to some 'general toxicity' because specific changes in gene expression were observed (X.-P.L. and M.P., unpublished results) using the differential display technique²⁷. Thus, retinoid anti-proliferative activity can be separated from retinoid differentiation-inducing activity. The anti-AP-1-selective retinoids are of particular interest because of their anti-proliferative activity. Because they cannot induce transcriptional activation from RAREs, they are likely to have fewer side-effects and could be starting points for a new generation of retinoid therapeutic agents. □

- Morris-Kay, G. (ed.). *Retinoids in Normal Development and Teratogenesis* (Oxford Science, Oxford, 1992).
- Bollag, W. & Holdener, E. E. *Ann. Oncol.* **3**, 513–526 (1994).
- Kraemer, K. H. et al. *New Engl. J. Med.* **318**, 1633–1637 (1988).
- Smith, M. A. et al. *J. clin. Oncol.* **10**, 839–864 (1992).
- Huang, M. et al. *Blood* **72**, 567–572 (1988).
- Hong, W. K. et al. *New Engl. J. Med.* **323**, 795–801 (1990).
- Leid, M., Kastner, P. & Chambon, P. *Trends biochem. Sci.* **17**, 427–433 (1992).
- Zhang, X.-k. & Pfahl, M. *Trends Endocrin. Metab.* **4**, 10–16 (1993).
- Pfahl, M. *Semin. Cell Biol.* **5**, 95–103 (1994).
- Yang-Yen, H. F. et al. *New Biol.* **3**, 1206–1219 (1991).
- Schüle, R. et al. *Proc. natn. Acad. Sci. U.S.A.* **88**, 6092–6096 (1991).
- Salbert, G. et al. *Molec. Endocrin.* **7**, 1347–1356 (1993).
- Pfahl, M. *Endocr. Rev.* **14**, 651–658 (1993).
- Angel, P. & Karin, M. *Biochim. biophys. Acta* **1072**, 129–157 (1991).
- Lehmann, J. M. et al. *Cancer Res.* **51**, 4804–4809 (1991).
- Graupner, G. et al. *Biochem. biophys. Res. Commun.* **179**, 1554–1561 (1991).
- Delescluse, C. et al. *Molec. Pharmacol.* **40**, 556–562 (1991).
- Lehmann, J. M. et al. *Science* **258**, 1944–1946 (1992).
- Dawson, M. I. & Okamura, W. H. (eds) *Chemistry and Biology of Synthetic Retinoids* (CRC, Boca Raton, 1990).
- Lee, M.-O., Dawson, M. I. & Pfahl, M. *Proc. natn. Acad. Sci. U.S.A.* **91**, 5632–5636 (1994).
- Apfel, C. et al. *Proc. natn. Acad. Sci. U.S.A.* **89**, 7129–7133 (1992).
- Titcomb, M. S. et al. *Molec. Endocrin.* **8**, 870–877 (1994).
- Hu, L. & Gudas, L. J. *Molec. cell. Biol.* **10**, 391–396 (1990).
- Vasios, G. et al. *EMBO J.* **10**, 1149–1158 (1991).
- Liang, P. & Pardee, A. B. *Science* **257**, 967–971 (1992).
- Hussmann, M. et al. *Molec. cell. Biol.* **11**, 4097–4103 (1991).

Received 3 June; accepted 3 October 1994.

- Lotan, R. *Biochim. Biophys. Acta* **605**, 33–91 (1981).
- Roberts, A. B. & Sporn, M. B. in *The Retinoids* (eds Sporn, M. B., Roberts, A. B. & Goodman, D. S.) 209–286 (Academic, Orlando, 1984).

ACKNOWLEDGEMENTS. Supported by grants from the National Cancer Institute (M.P., M.I.D. and M.P.) and the University of California Tobacco-Related Disease Research Program (M.P.).

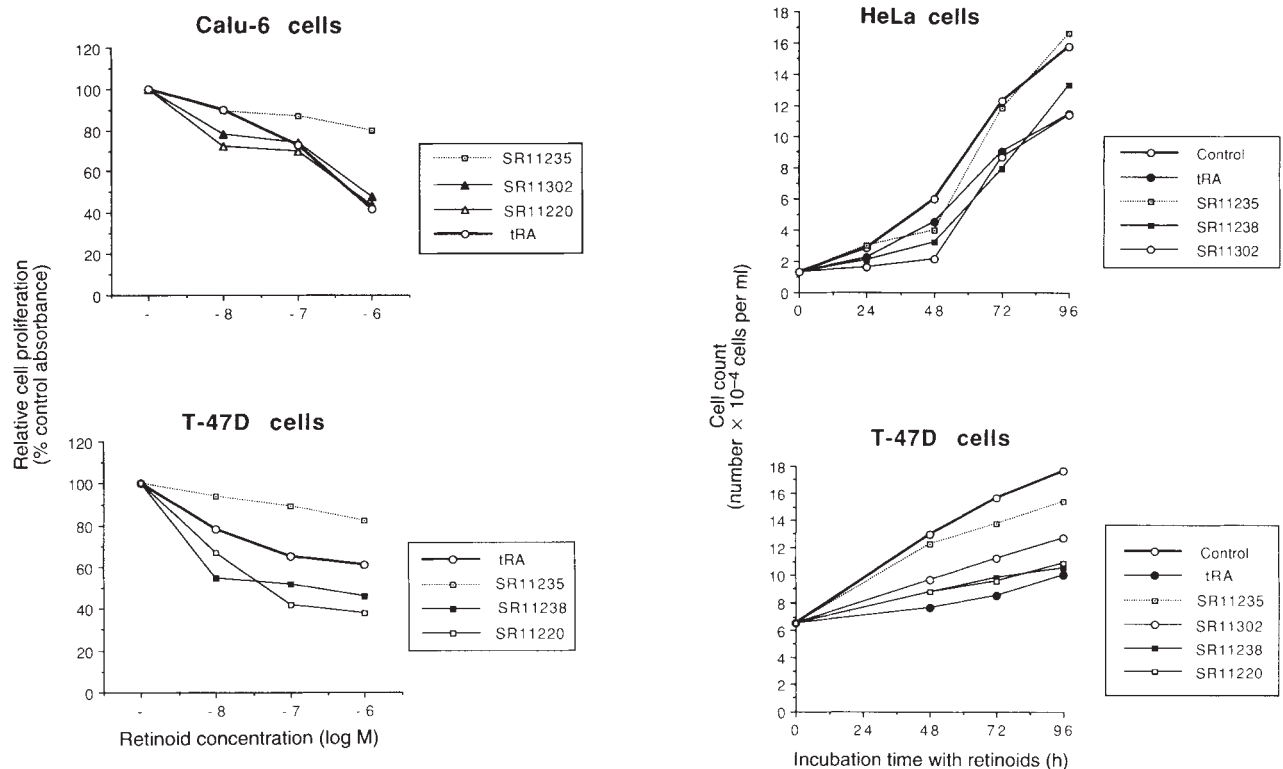


FIG. 3 Growth inhibition of cancer cells by retinoids. Proliferation of cancer cell lines was determined by using a colorimetric assay as described in Fig. 2 (Calu-6, T-47D) and/or by cell counting (T-47D, HeLa). Both procedures gave essentially the same results. For the cell proliferation kit assay Calu-6 human lung cancer cells and T-47D human breast cancer cells were seeded at 2,000–3,000 cells per well in 96-well plates (Costar). After 24 h, the cells were treated with various

concentrations of the indicated retinoids for 4 days and assayed for their capability to reduce MTT dye to a coloured formazan product as an index of cell survival and proliferation. The results are expressed as a percentage of absorbance at 550 nm of MTT-derived formazan developed by cells treated with control solvent. The growth of HeLa and T-47D cells in the absence and presence of retinoids (10^{-6} M) was also measured by cell counting.

Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix

Heinz W. Pley, Kevin M. Flaherty
& David B. McKay*

Beckman Laboratories for Structural Biology,
Department of Structural Biology, Sherman Fairchild Bldg,
Stanford University School of Medicine, Stanford,
California 94305–5400, USA

IN large structured RNAs, RNA hairpins in which the strands of the duplex stem are connected by a tetraloop of the consensus sequence 5'-GNRA (where N is any nucleotide, and R is either G or A) are unusually frequent¹. In group I introns there is a covariation in sequence between nucleotides in the third and fourth positions of the loop with specific distant base pairs in putative RNA duplex stems²: GNAA loops correlate with successive

5'-C-C · G-G base pairs in stems, whereas GNRA loops correlate with 5'-C-U · G-A. This has led to the suggestion that GNRA tetraloops may be involved in specific long-range tertiary interactions, with each A in position 3 or 4 of the loop interacting with a C-G base pair in the duplex, and G in position 3 interacting with a U-A base pair. This idea is supported experimentally for the GAAA loop of the P5b extension of the group I intron of *Tetrahymena thermophila*³ and the L9 GUGA terminal loop of the *td* intron of bacteriophage T4 (ref. 4). NMR has revealed the overall structure of the tetraloop for 12-nucleotide hairpins with GCAA and GAAA loops⁵ and models have been proposed for the interaction of GNRA tetraloops with base pairs in the minor groove of A-form RNA^{2,4}. Here we describe the crystal structure of an intermolecular complex between a GAAA tetraloop and an RNA helix. The interactions we observe correlate with the specificity of GNRA tetraloops inferred from phylogenetic studies, suggesting that this complex is a legitimate model for intramolecular tertiary interactions mediated by GNRA tetraloops in large structured RNAs.

The RNA sequence chosen for the structure determination of the hammerhead ribozyme has a GAAA tetraloop connecting the base-paired strands of stem II (see accompanying manuscript⁶). The crystals in which the structure was solved has three copies of the hammerhead in the crystallographic asymmetric unit. Fortunately, the crystal packing is such that the tetraloop of molecule 1 binds to the minor groove of stem II of molecule 2 (Fig. 1a); the tetraloops of the other two molecules

* To whom correspondence should be addressed.

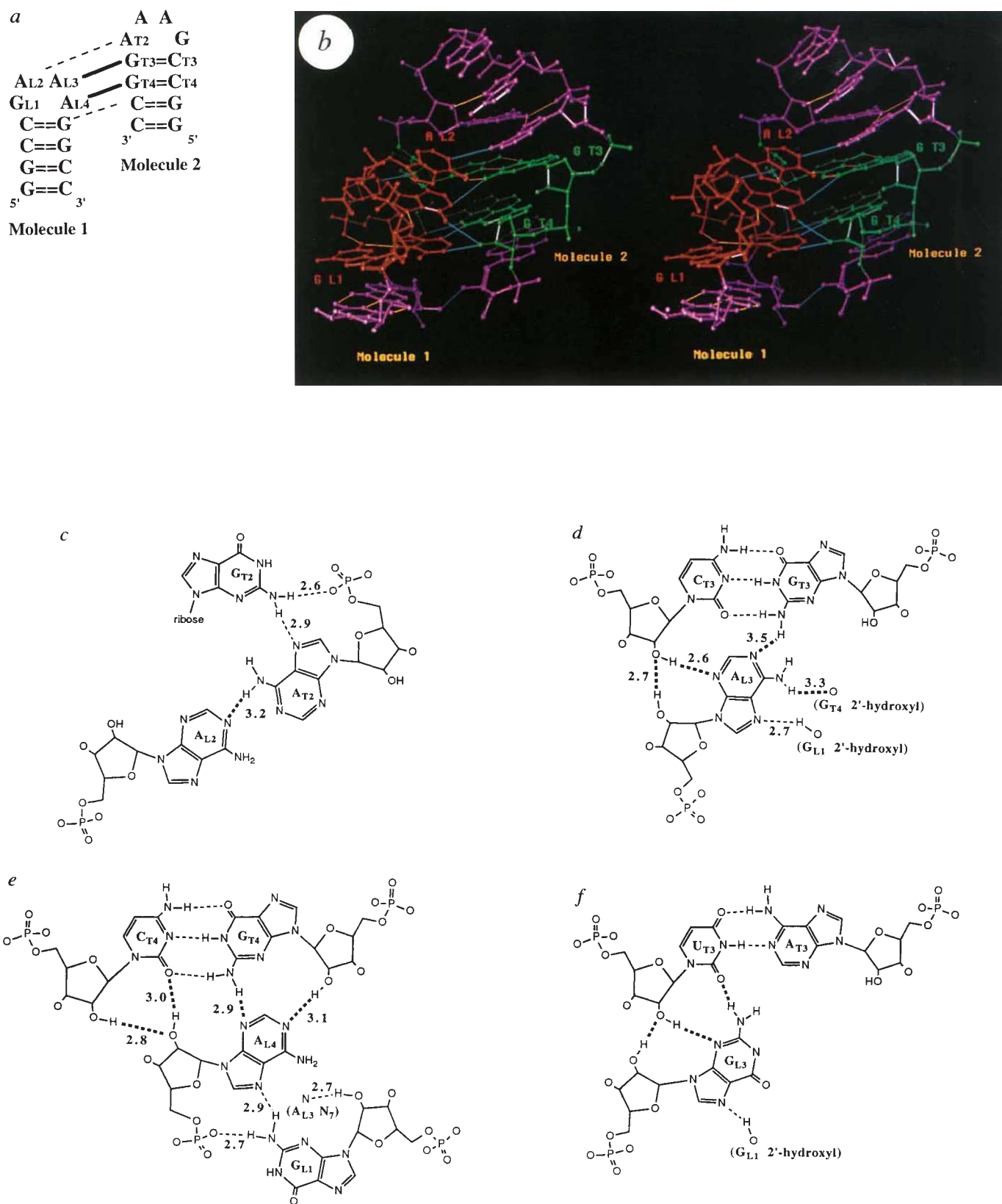


FIG. 1 Interaction of the 5'-GAAA tetraloop with an RNA stem. Nucleotides in the tetraloop are numbered successively L1, L2, L3 and L4. Target nucleotides in the RNA stem are numbered: T4 for the base pair interacting with A_{L4}, T3 for the base pair interacting with A_{L3}, and T2 for the bases that interact with A_{L2}. **a**, Schematic diagram of the interaction between the tetraloop of molecule 1 and stem II of molecule 2. **b**, Stereo model of the crystallographic structure. Tetraloop nucleotides are red; C-G base pairs in the T3 and T4 layers, green; other nucleotide,

magenta; intramolecular hydrogen bonds, yellow; intermolecular hydrogen bonds, cyan. **c-e**, Diagrams of interactions between nucleotides L2, L3 and L4 of the tetraloop and nucleotides in the RNA stem. Intramolecular hydrogen bonds are shown as thin dashed lines and intermolecular hydrogen bonds as thick dashed lines; distances between electronegative atoms (nitrogen or oxygen) of the bonds are in ångströms. **f**, Model showing how a guanosine at position L3 of the tetraloop might interact with a U-A base pair.

are not involved in crystal contacts. We therefore observe the structure of the GAAA loop in both 'bound' and 'unbound' states. Although the bound complex is intermolecular rather than intramolecular, it reveals at least one way in which a tetraloop of this class specifically recognizes base pairs in the minor groove of an A-form RNA helix.

The overall structure of the tetraloop is similar to that determined by NMR⁵. All of the nucleotides are in *anti* conformation, and the three adenine bases form a continuous stack with the bases on the 3' side of the helical stem. The three tertiary interactions we observe within the tetraloop are hydrogen bonds from the exocyclic amino group of G_{L1} (see Fig. 1 legend for nomenclature) to both the N7 nitrogen and the *pro*-Rp non-bonded phosphate oxygen of A_{L4}, and hydrogen-bond from the 2'-hydroxyl of G_{L1} to the N7 nitrogen of A_{L3}. These interactions would restrict the tetraloop to a GNRR consensus sequence, because base-specific interactions involve the exocyclic amino group of G_{L1} and the N7 positions of the purines at positions 3 and 4 of the loop. Presumably the additional restriction on the consensus that position 4 must be an adenosine is dictated by interactions with the target RNA helix. A summary of bond lengths of interest for the three copies of the tetraloop is given in Table 1; there are no major differences in structure between the 'bound' and 'unbound' tetraloops.

The structure of the tetraloop-stem complex is shown in Fig. 1. The axes of the tetraloop stem and the target helix differ in direction by 31°. The majority of the interactions between the tetraloop and the minor groove of the RNA helix are in two layers, in which A_{L3} and A_{L4} interact with successive C-G base pairs. A_{L4} is nearly coplanar with its target C-G base pair and forms four hydrogen bonds to it. Notably, only one of the hydrogen bonds (from the exocyclic amino group of G_{T4} to N3 of A_{L4}) is a base-base interaction; the other three involve 2'-hydroxyls of the sugars. A_{L3}, which has substantially greater tilt relative to its target C-G base pair, also has four potential hydrogen bonds. Again, only one is a base-base interaction; the other three involve ribose 2'-hydroxyls. Two of these interactions (from the exocyclic amino group of G_{T3} to N1 of A_{L3}, and from the 2'-hydroxyl of G_{T4} to the exocyclic amino group of A_{L3}) have distances significantly greater than 3 Å in the structure; they are likely to be relatively weak bonds.

In addition to these interactions, there is a single hydrogen bond between N1 of A_{L2} and the exocyclic amino group of an adenosine, and an intermolecular hydrogen bond between the 2'-hydroxyls of riboses in the duplex stems preceding the GAAA loop and its target. The former is a (presumably fortuitous) interaction with a nucleotide in the non-duplex GAAA loop of the target hairpin strand. The latter is an interaction between the helical segments of the RNA strands that does not involve the bases of either stem; this is likely to be a general, sequence-independent interaction that would be expected in other complexes between GNRA hairpins and A-form helices.

The interactions we observe in the GAAA-RNA stem complex correlate with the proposed specificity of tetraloops for sites on RNA duplexes observed by Michel and Westhof², although the way in which this specificity is achieved differs from that proposed^{2,4}. Transversion of the C-G base pair to G-C in either target layer would result in incompatible hydrogen-bond donor-acceptor pairing, thereby affecting one or two of the nine specific interactions between the tetraloop and its target. A transition to a target U-A base pair in either layer would delete a specific base-base hydrogen bond with an adenosine in the tetraloop by removing the exocyclic amino group of guanosine from the minor groove; however, it would presumably leave other interactions intact.

We have used our structure to model the interaction of a GAGA tetraloop with a target RNA stem. A direct replacement of A with G at position L3 of our structure results in steric clash of the exocyclic amino group of the guanosine base in the loop with O2 of the cytosine of the target base pair (with a distance of 1.8 Å between N2 of G_{L3} and O2 of C_{T3}); it also juxtaposes two hydrogen-bond donors, the two exocyclic amino groups of the guanine in the loop and target. Slight adjustments of G_{L3} of a model GAGA tetraloop bound to its cognate target base pairs, 5'-C-U · G-A, relieves the steric clash and allows a hydrogen bond from the exocyclic amino group of G_{L3} to O2 of U_{T3} (Fig. 1f). It is notable in this context that phylogenetic comparisons reveal substantial tolerance for base pairs other than C-G as the target of A_{L3} of GNAA tetraloops, but no tolerance of C-G or G-C and minimal tolerance of A-U as the target of G_{L3} of GNAA loops².

Many of the interactions we observe between the GAAA tetraloop and the RNA duplex helix include 2' hydroxyls of riboses, and would be unaltered by changes in sequence of the target duplex, so long as the L3 position of the tetraloop is G or A. Indeed, of the nine probable hydrogen bonds we observe (excluding the bond from A_{L2} to A_{T2} in the non-duplex loop of the target helix, which we presume to be specific to this complex), seven would be invariant if A were changed to G at position L3 of the tetraloop and C-G were changed to U-A as the target of A_{L4}. Thus, the primary determinant of sequence specificity in the interaction of GNRA tetraloops with RNA helices is expected to result from the combined effects of steric and hydrogen bond donor-donor clash of a G at the L3 site of the tetraloop with a C-G or G-C base pair in the RNA helix, rather than from specific base-base hydrogen bonding patterns between the tetraloop and its target.

The intermolecular complex we observe of a GAAA tetraloop with the minor groove of an RNA helix shows interactions that are consistent with the recognition specificity inferred from phylogenetic studies². The specific interactions we observe are likely to share many common features with, and may well be identical to, the intramolecular tertiary interactions that are thought to occur between GNRA loops and RNA helices in large, structured RNAs. □

TABLE 1 Interatomic distances between electronegative atoms within the tetraloops (Å)

	2-amino of G _{L1} -N7 of A _{L4}	2-amino of G _{L1} -phosphate oxygen of A _{L4}	2'-OH of G _{L1} -N7 of A _{L3}	N3 of G _{L1} -6-amino of A _{L4}
Molecule 1	2.9	2.7	2.7	3.9
Molecule 2	3.0	2.6	2.9	4.4
Molecule 3	2.9	2.6	2.7	4.1

Received 23 August; accepted 12 October 1994.

1. Woese, C. R., Winker, S. & Gutell, R. R. *Proc. natn. Acad. Sci. U.S.A.* **87**, 8467-8471 (1990).
2. Michel, F. & Westhof, E. *J. molec. Biol.* **216**, 585-610 (1990).
3. Murphy, F. L. & Cech, T. R. *J. molec. Biol.* **236**, 49-63 (1994).
4. Jaeger, L., Michel, F. & Westhof, E. *J. molec. Biol.* **236**, 1271-1276 (1994).
5. Heus, H. A. & Pardi, A. *Science* **253**, 191-194 (1991).
6. Piey, H. W., Flaherty, K. M. & McKay, D. B. *Nature* **372**, 68-74 (1994).

ACKNOWLEDGEMENTS. We thank O. Uhlenbeck for discussion. This work was supported by a grant to D.B.M. from the NIH and by the resources of the Beckman Laboratories for Structural Biology.

First-line defence

Structural insight into the trimeric organization of the human mannose-binding protein sheds light on the workings of the non-clonal, innate immune response in vertebrates.

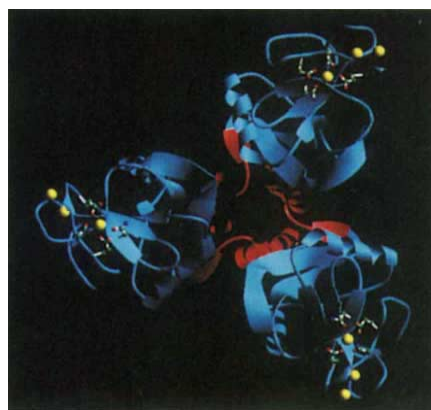
IN vertebrates, the highly specific antibody and cell-mediated immune responses can neutralize and eliminate invading pathogens with devastating precision. But because of the time taken (one to three days) to muster these responses, they cannot provide immediate protection: instead, the first line of defence against infection is an innate, or non-clonal, immunity that is activated within minutes. Invading organisms are marked for destruction as soon as they are spotted by a surveillance system of 'pattern-recognition' proteins, found for example on the surface of phagocytes in the circulation. The structure of a portion of one of these pattern-recognition proteins, human mannose-binding protein (MBP), is described in the latest issue of *Nature Structural Biology*¹, providing insight into the function of this primitive yet important arm of the innate immune system.

The discrimination between self and non-self provided by MBP and other members of the family of collectins² relies on their selective recognition of configurations of carbohydrates that are present on the cell walls of bacterial and fungal pathogens and on viral-surface glycoproteins, but which are only rarely present on mammalian cell-surface glycoproteins. MBP-decorated organisms are marked out for neutralization by complement-mediated cell lysis or opsonization; in the case of viruses, MBP can also block the interaction between virus and viral receptors on the host-cell surface.

The innate response is likely to be particularly important in children aged between six months and two years, when there is no longer any maternal antibody protection and whose immature immune systems cannot yet mount an efficient humoral response. Reduced levels of MBP, often due to mutations in the MBP gene, may be associated with an increased susceptibility to severe repeated bacterial infections in very young children.

MBP (M_r 32,000) consists of four domains: a short amino-terminal segment

rich in cysteine, followed by a collagen domain, which lies amino-terminal to a 40-amino-acid 'neck' region, and finally a carboxy-terminal carbohydrate-recognition domain (CRD). This overall subunit structure is shared by all the collectins².



View down the 3-fold axis of mannose-binding protein showing the carbohydrate-recognition domain (cyan) and neck region (red). Also shown are bound calcium ions (yellow). From ref. 1.

MBP is present in the serum as a hexamer of trimers, resembling the 'bunch of tulips' conformation of the first complement component, C1q. This multimeric form of MBP is essential for its function: whereas individual subunits have a low affinity for ligand, the 18 CRDs of the hexameric complex give a combined dissociation constant for mannan of 2×10^{-9} M (ref. 2 and references therein).

The isolated rat CRD, whose structure is known both in the absence and presence of ligand^{3,4}, forms non-physiological dimers in solution and in crystals. N-terminal addition of the neck domain adjacent to the CRD (Δ MBP) results in the formation of trimers in both solution and crystals. Sheriff and colleagues¹ have now determined the structure of trimeric Δ MBP. The three subunits are non-covalently joined by the helical neck regions, which together form a triple α -helical coiled coil. A similar triple-helical coiled-coil structure has been determined for a recombinant peptide of the neck region of another member of the collectins, lung surfactant D, indicating that the motif is probably a common structural feature of this protein family⁵. Furthermore, interactions between hydrophobic residues on the outside of the coil and hydrophobic residues of a neighbouring CRD also

contribute to the stability of the trimer¹.

The structure of the Δ MBP CRD is essentially identical with that of the isolated CRD (either with or without ligand), but the disposition of the CRDs is very different in the two structures. The orientation of the three carbohydrate-binding sites, positioned approximately 45 Å from each other, indicates that the subunits of the trimers recognize spatially distinct regions on the surface of the antigen, which, as the authors point out, provides for high avidity. Examining different crystal forms of the trimer reveals small differences in the relative orientation of the CRDs and neck domains. Sheriff and colleagues speculate that the CRDs have some freedom of movement relative to each other in solution and that this would increase the affinity of binding by providing cooperativity between CRDs. Further flexibility for carbohydrate binding is probably provided by the collagen helices that form the stalks of the hexameric bundle of tulips.

As well as forming the basis of trimerization, it is possible that the α -helical bundle serves as the nucleation point for the formation of the triple helix of the collagen domain. If the collagen domain is not involved in determining the trimer structure of MBP, what might its role be? It is likely that, for the collectins in general, this domain is important for binding to the C1q receptor on the complement-activation pathway. For MBP, it has also been suggested that the collagen domain functions as the binding site for the MBP-associated proteinase, also involved in complement activation. Consistent with this view, a mutant allele of MBP, with aspartic acid instead of glycine in the fifth collagen repeat, will form multimers but cannot activate complement¹. The structural features of the collectin family should between them help to expose the secrets of the vanguard defence operation.

Guy Riddihough

Guy Riddihough is Editor of *Nature Structural Biology*.

Also in this month's *Nature Structural Biology*: constructing the proteasome; determinants of G protein/effecter specificity; the interaction of a drug with calmodulin; snake toxin structure; an SH3-like protein and its interaction with DNA; the structure of the pleckstrin-homology domain from dynamin; and modelling RNA loop structures.

1. Sheriff, S., Chang, C. Y. Y. & Ezekowitz, R. A. B. *Nature struct. Biol.* **1**, 789–794 (1994).
2. Hoppe, H.-J. & Reid, K. B. M. *Prot. Sci.* **3**, 1143–1158 (1994).
3. Weis, W. I., Kahn, R., Fourme, R., Drickamer, K. & Hendrickson, W. A. *Science* **254**, 1608–1615 (1991).
4. Weis, W. I., Drickamer, K. & Hendrickson, W. A. *Nature* **360**, 127–134 (1992).
5. Hoppe, H.-J., Barlow, P. N. & Reid, K. B. M. *FEBS Lett.* **344**, 191–195 (1994).